EMERGING SOLUTIONS FOR MECHANIZING THE STORAGE AND RETRIEVAL OF INFORMATION

(Studies in Coordinate Indexing, Vol. V)

Compiled by
Mortimer Taube

DOCUMENTATION
INCORPORATED

1959

Printed in The United States of America

INTRODUCTION

While this volume was being prepared for press, we received two pamphlets from the National Science Foundation:

"Current Research and Development in Scientific Documentation, No. 5," October 1959

and

"Nonconventional Technical Information Systems in Current Use, No. 2," September 1959.

These pamphlets certainly indicate a great deal of activity in this field and it is out of this total activity that we have selected a number of our own papers and papers emanating from other organizations which can seriously be characterized as "emerging solutions." This phrase in itself indicates that the goal is not yet reached.

We have made a sustained effort to examine the results of all activities in this field but it is still possible that there are important contributions which we have overlooked. There will be subsequent volumes in this series and we hope at that time to repair any important omissions which we ourselves can discover or which are brought to our attention by other organizations active in this field.

TABLE OF CONTENTS

CHAPTER I

PROBLEMS OF MECHANIZING STORAGE AND RETRIEVAL OF INFORMATION*

By Mortimer Taube

Devices for mechanizing the retrieval of information have one major characteristic in common with all other machines - they exist only to reduce the cost in human effort of performing various tasks and, often, to perform those tasks faster than can be done by unassisted human effort, which is another way of saying: at a reduced cost. When the characteristics of information retrieval machines are examined, it will be apparent that these characteristics represent not the abilities to perform tasks which are qualitatively different from those performed by existing devices or systems, but only the ability to perform the same tasks at different rates, i.e., at different costs. For example, anything which can be selected in one search of the store by the most highly developed internal logic of a large general-purpose computer, can be selected by a card sorter, which selects on a single column at a time, by making as many successive searches as are required by the complexity of the question.

In other words, once the functions of a storage and retrieval system have been set forth, any device to perform such

---

* This paper is a general and broad-brush treatment of the topics considered in more detail in other papers in this volume. Presented originally at A.I.Ch.E. Meeting, Philadelphia, Pennsylvania, June 1958.

functions will differ from any other in cost. There are no qualitative differences among the indexing or searching potentials of different devices and systems. By recognizing that the differences are differences in cost, it is not implied that such differences are not important. The aim of mechanization of storage and retrieval systems is to make possible the reduction of cost (in human time and capital investment) of storing and retrieving information. It remains as a task of theory to analyze the requirements and operating parameters which determine total cost, and to provide against decisions based on only part of the relevant data, e.g., search speed per bit in abstraction from input cost, programming cost, etc.

With this general background in mind, that machines provide no magical solutions to the problems of information retrieval but only faster and cheaper ways of performing tasks already solved in principle, let us examine first certain fundamental considerations concerning the relationship of storage to retrieval.

A storage and retrieval system in its simplest terms is an organized method for putting items away in a manner which permits or facilitates their recall or retrieval from storage. This definition, although essentially circular, is intended to establish that a storage and retrieval system must be considered as a single system and not as a storage system plus a retrieval system. It is unfortunate that we lack a single word which expresses the total complex. For example, a communication system includes transmitters and receivers, and the over-all requirements of a communication system will determine a compatibility or correlation of design and function between its transmitting aspect and its receiving aspect. But the lack of a single accepted term to describe a storage and retrieval system has led in much of the literature on the subject to separate considerations of the store and the retrieval apparatus. For example,

there is considerable literature on microreproduction of documentary material which treats the retrieval of an item from the store as an afterthought and not a basic design requirement of the store itself. Similarly, there is much material on systems and devices for retrieving information which fails to consider the effect of the logical and physical design of the store on the design and performance of the retrieval apparatus.

It is obvious that the simplest way to put something away is to pile it in a heap. The easiest way to fill a warehouse is to put material away as it comes in without regard for organization or order until the warehouse contains a solid cube of stored material. At the same time, it is also obvious that finding a particular item in such a warehouse might entail emptying the whole warehouse (as a woman empties her pocketbook) and physically handling and examining every item in it. At the other extreme we might design a warehouse so that every item in it was preloaded on a conveyor running through the warehouse, with no item in the path of any other. Such a warehouse would be complex and expensive, but it would make possible push-button retrieval of any item.

Thus, it becomes obvious that the method of retrieving information is inseparably bound to the method of storing that information, because how information is stored largely determines how one must attempt to retrieve it. This is true not only of mechanized systems but also of manual systems.

Matching the Store Against the Question

The selection of an item from a store, if performed directly by a human being, involves an act of recognition. The human being who searches a store does so with a question in his mind which he matches against the properties of the items

in the store. The degree of similarity which would constitute a match between properties of items in the store and the requirements of a question in the mind of a human searcher cannot be exactly specified. Any physical item can be characterized by an infinite set of properties, any subset of which must function as the basis of human recognition and selection.

The substitution of mechanical recognition for human recognition seems to depend (1) on the possibility of describing an item in the store with a finite set of properties; (2) on the possibility of using codes to represent such properties; and (3) on the possibility of specifying the conditions of a match or failure to match on an all-or-none basis. This last condition states the requirement for converting the human recognition that "this is similar to that" into the machine recognition that "this code is equivalent (or is not equivalent) to that code."

It then becomes appropriate to a discussion of information machines to discuss characteristics of data-processing systems which are relevant to problems of information storage and retrieval, for information machines are indisputably data-processing devices, of which there is an increasing and bewildering variety now available.

Within the general field of automatic data processing or data handling, we distinguish three types of systems involving three different types of automata: (1) data-computation systems and computing automata; (2) data-transmission systems and communication automata; and (3) data storage and retrieval systems and "look-up" automata. An integrated data-processing system can involve all three types of data systems; but for any particular requirement one or another design characteristic is usually most important. A communication network can employ

computer-like devices as adjuncts for coding and decoding; a computer can have a limited look-up capacity as an internal store; and a storage and retrieval device can have teletype output (communication) and computer-type comparators. Even if an operating system exhibited characteristics of all three systems, it would still be necessary to distinguish the three types in the same sense that we distinguish a conductor as capacitor, resistor, or inductor even though it always has the properties of all three.

There are five characteristics of data-processing equipment which seem to have direct relevance to the information storage and retrieval problem:

1. Dense packing of codes which reduces the size of the store.

2. Rapid matching of the store against the question which reduces the time of search.

3. The "erasability" of the store which permits updating and elimination of obsolete information.

4. Matching many-termed questions against the store in one search, i.e., an increase in the degree of parallel access as compared with ordinary punched-card equipment.

5. The ability to program a single search to select on the basis of products, sums, and complements of classes as well as temporal order.

Since in storage and retrieval systems the size of the store and time of search seem to be of major importance, devices which have the characteristics in 1 and 2 must be accorded serious attention. On the other hand "erasability" and the ability to update may be important only for rapidly changing systems and may constitute a disvalue for

systems requiring permanent storage or even long-term storage. Characteristics 4 and 5 are obviously relevant to rate of search, i.e., by increasing the number of code elements read in parallel, the time of search can be reduced. In other words, the time of search for any given sized store can be reduced by packing codes denser and denser and moving such codes past a reading head faster and faster. Contrariwise, the packing density and rate of transport can be held constant and search time can be reduced by providing more reading heads.

Effect of Environment Upon Information Storage and Retrieval Systems

The environment in which an information storage and retrieval system operates has significant effects upon the system. The environment of a storage and retrieval system is the totality of those conditions which individually and collectively establish the requirements which must be met by the system. In a discussion of these conditions emphasis shifts from input to required rate of input; from searching or logical operations to required rate of searching or operating; from output to rate and form of output; from the logic of matching a question against the store to the size of the question and the number of questions which must be matched against the store in a given period of time; from maximum efficiency of coding to acceptable noise ratio; from the distinction between item codes and term codes to the determination of the absolute and relative sizes of the collection of item codes and term codes, etc.

A description of the environment states the requirements which must be met "efficiently" by a system or a device. These requirements can still be considered in theoretical or general terms; but some consideration of them seems to be necessary to establish the design requirements of any

system or device. It is not likely that any particular storage and retrieval device will be the "best" for all kinds of environment. But this does not mean that the choice of a particular device or system for a particular environment is arbitrary. Rather, the choice must be carefully considered, for the stakes in the game are high.

Terms and Items

In any storage and retrieval system, we deal ultimately with two kinds of entities: terms and items. An item is the thing we put away, a report, an abstract, a personnel file, a piece of hardware, a patent, an engineering drawing, a specification, etc. A term is a name, description, classification, numerical value or in general, any discriminator by which we characterize an item so that we can recall or retrieve desired items from a store. If each item in the store had only one characterizing term, like a number or like a name in the telephone book, retrieving a desired item from the store would be a relatively simple business - we could go immediately to a fixed position in the store to find any item. But when an item is characterized by a set of terms, any one of which or any combination of which may be used as a retrieval code, then it becomes impossible to locate the item in a fixed array by any of its terms. This creates the requirement for a searching apparatus which will search for an item characterized by a set of codes and reveal the address of that item.

Since there are only two entities, terms and items, there are only two ways to group codes in a storage and retrieval system. We can make a physical record for each item and enter the applicable term codes on that record (conventional grouping); or we can make a physical record for each term code and enter the applicable item codes on that record (inverted grouping). Logically the two systems are identical as indicated by Figure 1(a) and (b).

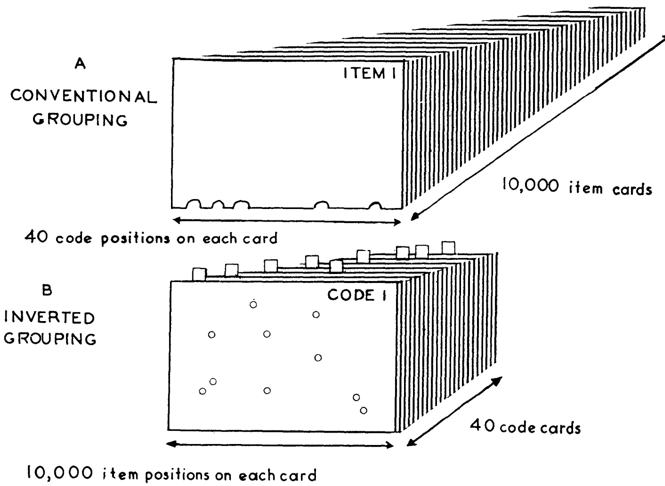


Figure 1 (a) and (b) Conventional and inverted methods of grouping codes in a retrieval system. The former requires a search of total file; the latter involves comparison of item codes on a selected set of term cards.

But there is an enormous difference in the efficiency or search times of the two methods. Conventional grouping of codes requires that a search be a sequential examination of the total file; but inverted grouping involves only the selection and comparison of item codes on a selected set of term cards, namely those which match the terms of a question put to the system.

This situation can also be illustrated with reference to mechanized searching by comparing punched-card sorting systems with punched-card collating systems.

Searching Systems

Searching is performed with a sorter by making several successive sorts until all items coded by a certain term or terms have been selected, that is, sorted out from the rest of the deck. Changing the column selector in the sorter

and selecting the cards from the proper pocket constitute setting up a question in the reading head of a piece of apparatus, and this question is matched successively against codes in the store. It is assumed that each item is represented by a card or set of cards on which is grouped the term codes characterizing that item.

With standard punched-card equipment (unless superimposed punching or wiring is used), the term codes in the question must be matched against specified fields on the item cards. For example, a simple sorter "sorts" cards column by column as determined by setting the column selector in the sorter; and the IBM "101" machine, which can search many columns at once, must be programmed to search in the proper columns for term codes in the question. The "101" can be wired to search for a single code in any of a number of fields. In such a system the total file must be sorted for each search.

Collating Systems

The inherent inefficiency of linear search has so far precluded the successful application of punched-card searching to collections of any significant size which cannot be divided into mutually exclusive classes; but several relatively successful punched-card installations have been organized for collating rather than searching. In setting up a system of punched cards for collating as contrasted with searching, grouping of items by terms is employed rather than grouping of terms by items. Table 1, in which the numbers indicate items and the letters indicate terms, illustrates the two forms of grouping.

When collation is used as a matching technique, item codes collected under one term are matched against item codes collected under another term. In effect one group of item codes becomes the question which is matched

Table 1.

Searching	Collating									
1 A M N O	A	1		3						
2 B C D T	B		2	3						9
3 A B M R	C		2			5				9
4 L N O P	D		2							
5 C G H K	F						6			
6 F G M P	G					5	6		8	
7 L P R T	H					5				
8 H K L S	K					5			8	
9 B C R S	L				4			7	8	
etc.	M	1		3			6			
	N	1			4					
	O	1			4					
	P				4		6	7		
	R			3				7		9
	S								8	9
	T		2					7		

against the other group considered as the store. It should be apparent that collation does not require the search of the total store but only of those item codes grouped under the terms of the question.

However, with standard collating equipment a considerable price must be paid for this decrease in search time. A collection of 1,000,000 items indexed by an average of 20 terms would require a file of 20,000,000 cards. With 10,000 terms in the vocabulary the 20,000,000 cards would be arranged in 10,000 groups averaging 2,000 cards in a group. Since a standard collator feeds 240 cards per minute from each feed, the collation of two terms (asking a two-termed question) would average between 10 and 20 minutes. This is an appreciable reduction from the time required for a sequential search of the total file but there

are some penalties which must be faced which reduce radically the efficiency of standard collators as information searching devices.

First, the size of the store must be increased enormously to permit prefiling items (cards) under every term by which they are indexed, in this instance, from 1,000,000 to 20,000,000. Second, collators only work on arrays maintained in fixed numerical or alphabetical order. Hence item cards must be filed (posted) to each term array and maintained in that array in a fixed order. Third, cards matched by the collator and selected as answers must be refiled in proper order. If the selected cards are to be retained as an answer or are to be matched against other groups, they may have to be duplicated so that the array from which they are selected initially can be restored to completeness for other searches.

Magnetic-tape Systems

In an examination of magnetic tape systems, of course, one immediately considers computers. One of the difficulties of evaluating any general-purpose computer in the abstract for information storage and retrieval is that such devices are extremely flexible and are made up of many different components. A total system may employ cores, drums, discs, tapes, and punched cards, and there may be more of certain elements than others depending upon the requirements the computer is literally "put together" to satisfy. Any organization which has available to it a general-purpose computer may use some of its components and subsystems for the storage and retrieval of information. This creates the problem of determining what percentage, if any, of a computer's cost should be charged against the storage and retrieval system, or even what characteristics or capacities of a computer are relevant to this specialized use as opposed

to other uses served by a computer.

Magnetic-tape systems are linear systems which provide rapid transport of tape past a reading head and high density storage of code elements. Magnetized dots on the tape can be packed 500 to the inch and multiple channels and reading heads can be provided for various widths of tape. The width of the tape and the number of reading heads determine the number of code elements which can be read in parallel.

Although most experiments with magnetic-tape devices have grouped term codes under item codes with the groups randomly arranged on the tape, the high packing density and high transport speeds of tapes have encouraged experiments in both inverted grouping of codes (item codes under term codes) and in multiple entry of item groups. In other words, it is possible to use tapes as a linear searching system or as a collating system, just as with punched cards. In addition, searching time can be reduced by entering an item and its term codes in several different places on a tape or on several different tapes, just as many cards for an item are filed in a catalog. This method of multiple filing on tapes, while difficult, is not impossible. However, multiple filing and multiple access are best provided with systems using discrete elements like punched cards.

It is necessary to emphasize that the inefficiency of sequential searching is a matter of principle because there are those who hope to overcome this inefficiency by spending more and more dollars for data-processing equipment with faster and faster rates of search. The situation here has been well summed up by R. A. Fairthorne:

"The rate and cost of access to a required item is proportional to a fractional power of the number of items to be searched. With linear access, such as single speed tape or film searching on a single reel, it is directly

proportional to the number of items. Therefore however slow multi-level (multi-dimensional) access may be with a small collection, and however fast linear scanning, there will be a certain size of collection over which the multi-level access will always be the faster. This is beginning to dawn on the engineers, who are now graduating from the ribbon to the scroll book. In due course they will triumphantly announce their rediscovery of the bound volume of pages"[1].

Conclusion

The fact that a computer can be used as a storage and retrieval device if considerations of efficiency are disregarded, does not establish computers as universal information-handling machines any more than the self-propelling property of steam shovels makes them universal vehicles of locomotion. Machines are designed for special purposes; the design and logic of any individual machine should reflect such purposes. If a complete abstraction is made from purpose and efficiency, there remains no basis for design, that is, no basis for the logical and physical arrangement of parts and functions which constitute a machine. Hence the concept of a universal machine is in essence contradictory.

Undoubtedly a great deal of money and ingenuity has gone into the task of investigating the adequacy of accounting machines, statistical machines, computing machines, and the like for the storage and retrieval of information. Even though this effort has not resulted in successfully operating mechanized systems, it has not been wholly or even largely useless. From it we have learned the design requirements of special-purpose storage and retrieval devices. At least one of these devices is now in the development stage. It promises to give access in minutes to million-item stores even when the items are characterized by as many as twenty terms. Furthermore

this device and others like it are not required to store elaborate programs and carry out complicated sets of instructions. They are matching machines and promise to be relatively inexpensive and simple to operate. Thus it can be expected that the storage and retrieval of information will yield rapidly to effective mechanization once the logic of the operation is understood, and the design of the hardware we develop is determined by that logic.

References

1. Fairthorne, R. A, "Matching of Operational Languages in Documentary Systems." Advisory Group for Aeronautical Research and Development Report No.49, NATO, Paris, p. 7 (1956).

CHAPTER II

THE DISTINCTION BETWEEN THE LOGIC OF COMPUTERS AND THE LOGIC OF STORAGE AND RETRIEVAL DEVICES*

By Mortimer Taube

This paper seeks to differentiate storage and retrieval devices from computers within the general class of automatic data handling devices. The basis for this differentiation is both negative and positive; negative in that the identity of computers and computer logic with automata and the logic of automata in general is shown to be in error; and positive in that the specific logic of storage and retrieval automata is shown to be the algebra of classes as distinct from the propositional calculus which is the logical instrument for the analysis of computers and computer circuits. As a final point, the time sequence in computer operation is shown to be analogous to non-commutative pairs in the algebra of classes.

Within the general field of automatic data processing or data handling, we distinguish three types of systems involving three different types of automata: (1) data computation systems and computing automata; (2) data transmission systems and communication automata; and (3) data

---

* Reprinted from the report submitted under the same title to the Information Complexes Division, Directorate of Mathematical Sciences, Air Force Office of Scientific Research, Contract No. AF 49(638)-91, September 1957.

storage and retrieval systems and "look-up" automata. An integrated data processing system can involve all three types of data systems; but for any particular requirement one or another design characteristic is usually most important. A communication network can employ computer-like devices as adjuncts for coding and decoding; a computer can have a limited look-up capacity as an internal store; and a storage and retrieval device can have teletype output (communication) and computer-type comparators. Even if an operating system exhibited characteristics of all three systems, it would still be necessary to distinguish the three types in the same sense that we distinguish a conductor as capacitor, resistor, or inductor even though it always has the properties of all three.

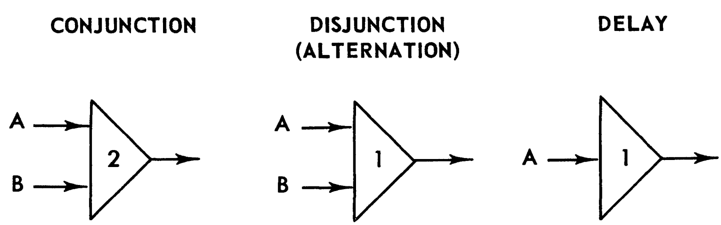
One of the significant implications of this distinction is that it restricts the notion of computers and computing automata; it presents computing automata as a special class of information handling devices and denies their universality.

The IRE Standards for Electronic Computers for 1956 gives the following definition of computers: "Computer. (1) A machine for carrying out calculations. (2) By extension, a machine for carrying out specified transformations on information." The first definition is accepted here and the second is rejected as ambiguous because it is obvious that there are many types of machines which are not computers and which carry out specified transformations on information, i.e., typewriters, telephones, cameras, phonographs, etc.

What seems to justify this extended definition is a growing literature[1] in which computers are identified with automata in general. The basic emphasis in much of this literature is the notion of a universal computer (or automaton) which may be either a Turing machine or a McCulloch-Pitts net.

Such devices are presented not only as general purpose computers but as general purpose automata which in extreme cases can do anything, even construct other devices to do what the first machine in the series cannot do.

The major theoretical basis for generalizing the notion of computers to encompass automata in general seems to lie in the supposition that it is possible to describe in logical terms a universal element or "basic organ", networks of which make up any possible automata. Thus Kleene[2] represents the basic organs of a net as a conjunctive net, a disjunctive net and a delay net. Although not carefully defined, negation is represented as the not-firing of an input. Von



Neumann ascribes to computing automata the basic organs A . B (A and B); A + B (A or B); and A^{-1} (not A).[3] Since these are the basic operations of the propositional calculus, Von Neumann notes that the three basic organs can be reduced to the Sheffer stroke function of joint denial.[4] Further, he points out that a computer can be described in terms of the propositional calculus plus time sequence. "These remarks enable us to describe the relationship between automata and the propositional calculus. Given a time delay S, there exists a one-to-one correspondence between single output automata with time delay S and the polynomials of the propositional calculus."[5]

It will be shown that the ability to describe computer networks in terms of the propositional calculus plus time delay indicates not their universality but their special character. The propositional calculus is not the whole of logic nor, from certain points of view, its most primitive and basic part. There are basic notions in logic, e.g., the notion of term, class and member which play no role in the propositional calculus. And, by the same token, there are devices, the analyses of which require other logical instruments. For example, the logic or algebra of classes is the branch of logic which is most appropriate for the analysis of storage and retrieval devices.

Actually the identification of computers with universal automata seems to be based on the mistaken assumption that the propositional calculus is the whole of logic. Hence, if the distinction between the propositional calculus and the logic of classes is clearly exhibited, this will serve to establish the special character of computers and the significance of information handling devices that are not computers. But before proceeding with this task it is necessary to give a brief description of what is meant by a storage and retrieval device as contrasted with a computer.

The simplest and best known storage and retrieval devices are printed indexes, or trays of cards. These devices always exhibit some arrangement which determines the requirements of "look-up" or retrieval. The arrangements, in turn can be characterized by various degrees of freedom ranging from a fixed numerical order to complete randomness. Between these two extremes there may be alphabetical arrangements, subject index entries, subject headings, arrangement by Uniterms, a hierarchial order of classes, or combinations of these arrangements. The important thing to note is that the "look-up" or retrieval requirement is a general characteristic of storage and retrieval devices and not something which depends upon a particular arrangement.

In addition to arrangement of information, every storage and retrieval system must provide for identification of (1) the item stored and (2) the characteristic by which the item is to be retrieved. The item may be a telephone number, a street address, a person, an abstract, a report, a book, a piece of hardware, etc. The characteristic can also take on many forms: a name, a number, a description, a code, a Uniterm, a subject heading, a class designation, etc.

The item must be stored so that a search by any characteristic will disclose either the address of the item, an abstract of the item, or the item itself. Just as computers can be built to replace hand calculators and desk calculators, so storage and retrieval devices can be built to replace card catalogs, printed indexes and dictionaries. The Minicard System is such a device; so are the Matrex machines and the Ediac; and the Peek-a-boo devices developed by the National Bureau of Standards. The Rapid Selector is a limiting case in a sense to be made clear below.

The distinction between item and characteristic is crucial in storage and retrieval systems. It indicates that the logical structure required by a storage and retrieval system is not the propositional calculus but the logic of classes. The characteristics are names of the classes and the items are members. This difference in logic is fundamental and establishes a basic distinction between computing automata and storage and retrieval automata.

The Propositional Calculus and the Logic of Classes[6]

In textbooks that are not scrupulous in their presentations, the massive analogies that exist between the propositional calculus and the logic (or algebra) of classes tend to obscure fundamental differences between these two branches of logic. Thus each of these algebras (of propositions and of classes) may be regarded as an interpretation of the abstract Boolean

algebra. But the two logics, of propositions and of classes, also have their very important differences. These differences stem from the fact that the notions of "element" and "member," which are fundamentally distinct for the logic of classes, can be thought of as coalescing for the logic of propositions.

Let us think of a "value" or an "element" as any constant to which the variables in an algebra (once it is interpreted) might be equated. Thus "element" and "value" will be used interchangeably in what follows.

The Boolean algebra postulated two such constants, elements or values, which are usually denoted by "0" and "1". <u>But it neither affirms nor denies that there exist any other values.</u>

By restricting the notion of element to the two values "0" and "1", the Boolean algebra can be interpreted as a propositional calculus in which "0" is equivalent to "falsity" and "1" is equivalent to "truth." This restricted interpretation of the Boolean algebra is explicitly noted by Shannon.[7] He presents the following postulates of switching circuits and notes they are "exactly analogous to the calculus of propositions":

1. a) $0 \cdot 0 = 0$
   b) $1 + 1 = 1$
2. a) $1 + 0 = 0 + 1 = 1$
   b) $0 \cdot 1 = 1 \cdot 0 = 0$
3. a) $0 + 0 = 0$
   b) $1 \cdot 1 = 1$
4. a) At any given time, $X = 0$ or $X = 1$

Shannon continues:

"The symbols of Boolean algebra admit of two logical interpretations. If interpreted in terms of classes the variables are not limited to the two possible values 0 and 1. This

interpretation is known as the algebra of classes. If, however, the terms are taken to represent propositions, we have the calculus of propositions in which variables are limited to the values 0 and 1.... Usually the two subjects are developed simultaneously from the same set of postulates except for the addition in the case of the calculus of propositions of a postulate equivalent to postulate 4 above."[8]

Both to illustrate how the logic of propositions and the logic of classes are two interpretations of the abstract Boolean algebra, and to keep discussions of the interpretations distinct from one another and from that of the abstract system, let us agree to the following symbolic transformations:

Boolean Algebra	Algebra of Propositions	Algebra of Classes
$x \cdot y$	$p \cdot q$	$X \cap Y$
$x + y$	$p \vee q$	$X \cup Y$
$- x$	$\sim p$	$\overline{X}$
0	F	Λ
1	T	$\mathbf{V}$

We may summarize by saying that both the algebra of propositions and the algebra of classes conform to every law of the Boolean algebra. In addition, the algebra of propositions satisfies Shannon's Postulate 4, quoted above. i. e.,

$$p \equiv T \text{ or } p \equiv F$$

Note: We write "$p \equiv T$" rather than "$p = T$" for two reasons: it uses the customary symbol of the usual algebra of propositions, and it emphasizes the fact that we are dealing with a

truth functional, rather than, say, a modal, algebra of propositions.

But this additional postulate has usually no counterpart in the algebra of classes. This is where the question of members comes in. Let us see what the situation is.

How many individuals are there having membership in the universal class, **V** ? The Boolean algebra insists that there is at least one; otherwise the universal class and the null class would not be distinct, as postulated in the Boolean algebra. But beyond this, it is impossible to say: it depends on what is the universe of discourse.

Let us assume--a thing not usually assumed in the algebra of classes, but still consistent with that algebra--that this number of individuals is finite. (This assumption is appropriate to the construction of storage and retrieval machines, since the universal class can be thought of as the class of all stored items, any subset of which is to be retrieved.) Let us designate this number by "n." In terms of it, a postulate analogous to Shannon's Postulate 4, above, may be formulated. Indeed Shannon's Postulate 4 follows immediately from the assumption that n = 1. (Here "1" is the cardinal number, not the Boolean constant.) LAW OF VALUES: If there are n members of **V** , then there are 2^n distinct values available for the variables "X," "Y," "Z," etc.

It may be seen how Shannon's Postulate 4 follows from the above law of values. Where n = 1, there are two distinct values for the variables in question. Since the Boolean algebra postulates two constants, 0 and 1, these two values must be identified with those two constants.

But what is this one member of the "universal class," "T" in the algebra of propositions? (Von Neumann suggests that we may take "T" or "1" as short for "pv~ p" and "F"

or "0" as short for "p.∼p" where p is some (unspecified) sentential constant. Thus "T" and "F" are propositions, or elements, or values, in the algebra of propositions.) We may call it truth, or the actual state of affairs. From one point of view, there is only one actual state of affairs, namely, the universe as it is, hence there is only one member of the universal class "T". Different propositions either describe this universe, and are equivalent to T, or fail to do so and are equivalent to F. The calculus of propositions does not discriminate propositions with respect to their meaning, or the aspect of truth of the universe that they report; but only with respect to their truth or falsity.

It is here that the algebra of propositions, so to speak, confuses element or value, with member. For we can take truth to be either the one and only member of 1, or to be that element 1 itself. This is because the universal class is a unit class.

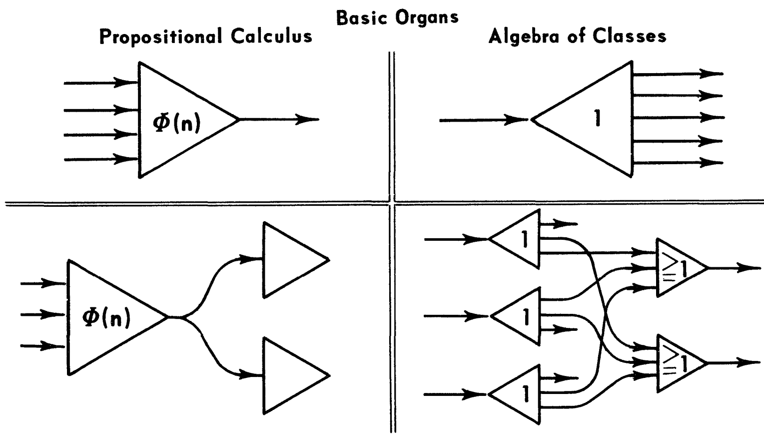
But most important applications of the algebra of classes will be ones in which the universal class is not a unit class, that is where $n>1$. In these cases, the number of distinct classes, that is, the number of values or elements in the algebra of classes, is going to be larger than 2. If n is very much larger than 1, the number of values will be vastly larger than 2, by the law of values.

In computing, the outcome is a report, true or false. But in storage and retrieval systems, the interest is in finding some determinate selection of the universal class. This selection may be all of them, or none of them or any intermediate aggregate of them.

Circuitwise this distinction is represented by the requirement that the basic organs of computing networks which operate according to the propositional calculus may have any

number of inputs but must have only one output. (Von Neumann[9], Kleene[10], Minsky[11], Burks[12]). This output may be stimulated (T or 1) or unstimulated (F or 0). Another distinction is the requirement that computing outputs can be split but not merged nor intersected.

On the other hand the basic organ of the algebra of classes has one input and a finite number of outputs; and products are made by merging outputs.



It is true that the physical representation of a product of two classes may be a set of "and" gates. This is equivalent to enumerating in a set of propositions those items which make up the product of the two classes. But the product of two classes is a class and not a proposition; physically multiple, it is the class of "and" circuits required to indicate the members common to the two classes. Only in the particular case of a class with only one member can a class be identified with its member. In such a case the identification of "and" with logical multiplication or "product" occasions no difficulty. But if a class has indeterminate membership ≥ 0 then it is incorrect to identify logical multiplication with logical conjunction, just as it is incorrect to

identify a set with one of its members.

On the other hand, it is always possible to dispense with the notion of set or class and the notion of membership in favor of the notion of "property." However, the simple statement $(z)(z \epsilon x \,.\, z \epsilon y)$ [for any z, z is a member of y and z is a member of x] must then be expanded into a set of propositions:

"z_1 has the property x . z_1 has the property y"

"z_2 has the property x . z_2 has the property y"

"z_3 has the property x . z_3 has the property y"

etc.

If the basic organs of a system represent classes, the addition of a new member involves adding an output to a pre-existing circuit, e.g., adding a new telephone outlet to a pre-existing trunk; adding a number to a pre-existing Uniterm Card; adding a card to the proper class of a pre-filed deck of punched cards; or drilling a hole in a Matrex Card. On the other hand, if the basic organs of computers are used, each new item added to the system requires a new set of circuits for all the propositions which describe it, e.g., a total circuit for each new telephone outlet because no trunk lines (classes) have been preestablished; a total code for the item and its properties on the next available segment of tape; or a total code for the item and its properties in a new punched card. This explains why computers which store and search for sets of propositions instead of for products of classes are so inefficient when used as storage and retrieval devices.

A Note on Commutativity in Computers and SR Devices

One of the basic differences between the propositional calculus and the logic of classes is found in the idea of commutativity. Commutativity presents one of the major problems in the design of both computers and storage and retrieval devices.

Both the propositional calculus and the Boolean Algebra contain theorems which state their commutativity:

$$A \,.\, B \equiv B \,.\, A \qquad\qquad x \frown y = y \frown x$$

$$A \vee B \equiv B \vee A \qquad\qquad x \smile y = y \smile x$$

But whereas the notion of membership permits defining unit classes and non-commutative pairs within the logic of classes[13], so that "x ; y" $\neq$ "y ; x," there is no analogous form within the propositional calculus. Here occurs the importance of the notion of a time series, mentioned above. If a delay circuit is added to the "and" and "or" circuits of a computer, non-commutativity is provided in the sense of before and after. Thus, to represent A as stimulated <u>before</u> B, a delay organ is placed between A and B. But such delay organs indicate, as Von Neumann is so careful to point out, that a computing automaton cannot be represented by the propositional calculus alone, but only by the propositional calculus plus time.

Whereas it makes sense to put a delay element between two statements so that $A_{(T_1)} \cdot B_{(T_2)}$ is distinguishable from $B_{(T_1)} \cdot A_{(T_2)}$, it makes no sense whatever to talk of the members of x being before the members of y in the product "$x \frown y$".

Boolean Algebra is by definition commutative. But Boolean Algebra does not exhaust the domain of the algebra of classes. Within this latter domain it is possible to define a non-commutative pair, "x ; y". The different pairs "x ; y" and "y ; x" will have different codes in a storage and retrieval device. It follows that a storage and retrieval device can be described completely within the logic of classes including a provision for relations, i.e., non-commutative pairs.

If, in an SR device, it is desirable to distinguish "man bites dog" from "dog bites man", they can be coded differently: ("x ; y", "y ; x"). This coding may involve a difference in symbols or it may involve a difference merely in the arrangement of symbols.

Whatever method is chosen to represent non-commutativity, it must be recognizable by the reading head. If different symbols or codes are used for "x ; y" and "y ; x", recognition of the difference is a single operation. On the other hand, if the same symbols are used and only their positions are used to designate non-commutativity, the reading head must possess enough circuitry or a memory element to register not only the symbols it reads but their relative positions.

If non-commutativity is important in a storage and retrieval system, as many have supposed, it can be provided in either of two ways which are logically identical, namely, by an increase in discrimination within the code or by an increase of discrimination in the reading head. Either method will involve an increase in costs which must be justified by a demonstration that commutative systems result in too high a noise level.

Conclusion

The fact that a computer can be used as a storage and retrieval device if considerations of efficiency are disregarded,

does not establish computers as universal information handling machines any more than the self-propelling property of steam-shovels makes them universal vehicles of locomotion. Machines are designed for special purposes; the design and logic of any individual machine should reflect such purposes. If a complete abstraction is made from purpose and efficiency, there remains no basis for design; that is, no basis for the logical and physical arrangement of parts and functions which constitutes a machine. Hence the concept of a universal machine is in essence contradictory.

Computers and storage and retrieval devices are different types of information handling machines. Having different purposes, they differ in design, operating characteristics and logic. The distinction which has been drawn between a two-valued propositional calculus and an algebra of classes, illustrates the fundamental character of these differences.

References

1. See especially Automata Studies, ed. by Shannon, C.E. and McCarthy, J. : Princeton, Princeton Univ. Press, 1956 (Annals of Mathematics, No. 34).

2. Kleene, S.C.: "Representation of Events in Nerve Nets and Finite Automata". In Automata Studies, p. 5.

3. Von Neumann, J.: "Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components." In Automata Studies, p. 47.

4. Ibid. p. 54.

5. Ibid. p. 48.

6. The author recognizes his great debt to Professor Henry Leonard of Michigan State University who criticized an earlier version of this paper and supplied the material on pages 19, 20, 21, 22, 23, and 24.

7. Shannon, C.E.: "A Symbolic Analysis of Switching Circuits." AIEE Transactions, v. 77, 1938, p. 713.

8. Ibid. p. 714.

9. "The network is subjected to one condition, however. Although the same output may be connected to several inputs, any one input is assumed to be connected to at most one output. It may be clearer to impose this restriction on the connecting lines by requiring that each input and each output be attached to exactly one line, to allow lines to be split into several lines but prohibit the merging of two or more lines." Von Neumann, J.: Op. Cit. p. 45-6.

10. "A nerve net is an arrangement of a finite number of neurons in which each endbulb of a neuron is adjacent to (infringes on) the soma of not more than one neuron." Kleene, S.C.: Op. Cit. p. 5.

11, 12. On this same point see Minsky, M.L.: "Some Universal Elements for Finite Automata," p. 118, in Automata Studies; and Burks, A.W. and Wright, J.B., "Theory of Logical Nets." IRE Proceedings, v. 41, No. 10, p. 1357.

13. "The notion of relation as a class of ordered pairs goes back to Pierce and the notation 'x;y' to Frege and Peano. But Weiner (1914) was the first to show that the ordered pair could be defined within the theory of classes."Quine, W.V.O., Mathematical Logic. Cambridge, Harvard Univ. Press, 1951, p. 201-202.

CHAPTER III

THE RELATION OF THE SIZE OF THE QUESTION TO THE WORK ACCOMPLISHED BY A STORAGE AND RETRIEVAL SYSTEM*

By Mortimer Taube and L.B. Heilprin**

This paper is concerned with the definition of work accomplished by a search of a storage and retrieval system. In a storage and retrieval system, the store is not transmitted but interrogated. The receiver plays an active role by presenting the question to the store. This intrusion of the "question" establishes a fundamental distinction between the notion of the amount of information transmitted which is basic in communication theory and work accomplished in a search which is the basic concept of storage and retrieval theory. In the mathematical analysis the unit of work accomplished is defined as the matching of one word in the question against one word in the store. The rate of work accomplished, or search power, is the number of units of work per unit of time. The bearing of this theoretical conclusion is discussed with special reference to two specific devices or systems, the Rapid Selector and the Minicard System.

---

*Reprinted from the report submitted under the same title to the Information Complexes Division, Directorate of Mathematical Sciences, Air Force Office of Scientific Research, Contract No. AF 49(638)-91, August 1957.

**Formerly of Documentation Incorporated, now with the Council on Library Resources, Washington, D. C.

The definitions of terms and items set forth in <u>Mathematical Foundations for a Storage and Retrieval Theory</u>[1] will be used here but certain additional definitions are necessary:

Element (of code): a code element is the smallest discriminable physical part of a code, e.g., a hole, notch, magnetic dot, dot on a film, electronic pulse, etc.. In defining search work an essential notion will be the time required to read a code element or set of code elements without reference to the arithmetical or letter value assigned to the element.

Coding Field: a spatial or temporal continuum in which the code elements are recorded. The sensing of an element involves not only its recognition as a mark (hole, dot, notch, etc.) but also its recognition as a member of a set of elements, each having a fixed position with reference to the continuum in which it is recorded.

Bit: a binary information storage unit. The word is an abbreviation of "binary digit" and refers to the information-storing value of a code element in a binary code consisting of two characters: 0 and 1, hole and no hole, pulse and no pulse, dot and dash; etc. The capacity in bits of a storage device is the logarithm to the base two of the number of possible states of the device. In binary coding, the bit is the storage capacity of each element of the code. But in other types of coding, e.g., decimal coding on a punched card, position on a Zato card or an Alpha Matrex card, each code element may have a value of many bits.

Character: a character is a letter, number, or similar symbol (a, b, 3, 7, *, !) represented by a code element or set of code elements. The representation is accomplished not only by the element or elements but by their order, that is

to say, their position in the coding field. Thus on an IBM punched card the hole punched in the bottom row represents "9". In a Morse code the temporal order of dots and dashes —• represents the letter N, whereas •— represents A.

Word: a word is a combination of one or more characters in a fixed spatial or temporal order. Therefore it is representable by a set of sets of code elements in fixed position with reference to the coding field.

Term: a term is a word used to describe an item for the purpose of storing it or retrieving it from a storage and retrieval system. Whether or not we restrict the concept of term to single words or allow it to signify words or phrases seems arbitrary. It is necessary to distinguish a term as a unit of description from a set of terms considered as a set of descriptions of an item.

Item: an item is a physical object; a document, record, book, map, patent, report, abstract, photograph, etc. which is the ultimate object of a search in a storage and retrieval system. In some systems, the physical object itself is stored and retrieved directly. In other systems the ultimate item is represented by an abstract, an address, or a set of terms which describe it. In this paper an item will be represented by a set of terms. To store an item will therefore mean to enter ("post") the code which represents the physical item, under each member of the set of its descriptive terms. This set is a subset of the total number of terms, or the index vocabulary.

Store: the store is the set of all subsets of terms physically recorded in some medium (cards, tapes, drums, wires, etc.).

Question: a question is the term or set of terms which is compared with the store in order to select those items from the store whose terms match the term or terms of the question.

It will be noted that whereas order enters into the definition of character and term it does not enter the definition of an item, a store, or a question. The mathematical significance of this distinction will be explored below. At this point it is only necessary to note that the order of elements in a character and the order of characters in a word in a store must be in one-to-one correspondence with the order of elements and characters in the question. This requirement does not exist with reference to the order of terms, although it is possible to require in any particular case that the terms in any question be matched against terms in an item in a certain order. We show below that any question limited to a fixed order of terms which is not necessarily matched by the order of terms in the store, will involve an increase in work accomplished or an increase in capacity of the internal memory. In other words it will cost money, electronics or time to ask a question in which the terms occur in a certain order, which order is not isomorphic with the order of terms in the store.

II

All the notions so far defined, except that of a question, have equal relevance for communication theory and storage and retrieval theory. The recipient of a communicated message is passive, that is, the communication system delivers to the recipient what is sent or transmitted by the sender. In a storage and retrieval system, the store is not transmitted but interrogated. The receiver plays an active role by presenting the question to the store. This intrusion of the question establishes a fundamental distinction between the notion of amount of information transmitted which

is basic in communication theory and work accomplished in a search which is the basic concept of storage and retrieval theory.

In communication theory, once the capacity of a channel per unit of time has been defined, a system is measured in terms of the amount of information it can transmit. In a noiseless channel the capacity of any channel determines the rate of transmission, such that Capacity x Time = Number of Messages. Since the capacity defines the rate at which information can be transmitted per unit of time, or rate of transmission, the formula R x T = D can also be applied in storage and retrieval theory. Let R be the rate of search, T the time of search and D the store. Up to this point there exists a parallelism between communication theory and storage and retrieval theory. The number of messages transmitted in a given time segment is analogous to the size of the store searched in a given time segment. But whereas the number of messages transmitted constitutes the work accomplished by a communications system, the store searched is not the measure of work accomplished by a storage and retrieval device. An additional variable is needed in the latter case, namely, a variable expressing the size and structure of the question.

The work accomplished in a search is the set of items in a store selected by matching the terms in the question against the terms recorded in a store to describe the items. One possible measure of the work accomplished would simply be the number of terms in the store which is read or examined. But the number of terms in the question is also a factor in work accomplished. Actually, the terms in the question and the terms in the store have identical mathematical relationships to the work accomplished. Consider a store of 1,000 terms and a question of three terms. Physically there are two methods of matching or searching. The terms in the question can be held stationary and the

store can be moved so that the 1,000 terms are sequentially matched against the three. Or the store can be held stationary and the three terms of the question can be moved over the store. So long as the object of the search is matching, and selection on the basis of matching, it is clear that the two procedures are mathematically equivalent. This means further that in relation to search work accomplished it does not matter whether we call the collection of three terms the question and the collection of 1,000 terms the store, or vice versa. *

This mathematical equivalence of question and store with reference to work accomplished has a practical consequence often overlooked by those who design storage and retrieval systems, namely, increasing the size of the question makes the same demands on the system as increasing the store to be searched. Searching a store of a thousand terms with a 3-term question is roughly equivalent to searching a store of 3,000 terms with a 1-term question, i.e., the store must be searched three times, or three times as many parallel matching circuits must be provided. ** If the question involves also the order of its terms, the store may have to be searched six times. The ability of a system to handle lengthy, involved questions may greatly increase the required search work. Therefore design requirements for search questions should be carefully limited to what is actually needed. Capacity for asking a complex question should not be included because of the mere assertion of its desirability. When the Minicard System is examined in this regard it will be seen that the problem of the size of the question turns out to be one of the major obstacles to

*In a Uniterm or Matrex System (Batten) each card in the system is part both of the store and of a question. The superimposition of one card over another can be considered a search of part of the store by a question.

**See Appendix.

the development of a successful system and presents one of the major parameters of cost.

III

In the mathematical analysis appended to this paper the unit of work accomplished is defined as the matching of one word in the question against one word in the store. The rate of work accomplished, or search power, is the number of units of work per unit of time. It should be noted that in addition to work accomplished by matching, any search of a storage and retrieval device involves memory, even if in the minimum case it is only the memory of which is the "select" and which is the "reject" receptacle.

Sequential search is search of a store using one term of a question at a time. Parallel search is simultaneous search of a store by a question with more than one term. Decreasing the time of search by increasing the rate of search in this manner requires increased circuitry and memory capacity beyond that of sequential search.

Now in a search of a store for one term of a question at a time, successive searches by additional terms involve searching a continually decreasing store. But parallel search involves searching the total store by all the terms in the question. This means that not only the rate of work but the amount of work which must be performed is increased in parallel search.

The point here can best be understood if we consider what occurs when we select a card or cards coded by three digits, say 387, from a deck of punched cards. The first sorting selects all cards coded XX7. Let us assume that one tenth of the cards are so coded. Then in the second sorting only one tenth of the original store must be examined, etc.. Hence the search work accomplished in suc-

cessive searches is reduced. But if it is desired to search for 387 in one pass of the cards, then the amount of comparison necessary to select each digit is constant and is a product of the number of cards times the number of digits (size of the question). This increased work is accomplished, as was stated, by incorporating in the selecting device increased electronic circuitry and memory capacity. Before considering the bearing of these facts on the design of selectors, two related matters must be introduced.

1. Organization of the Store.

Although there are many ways in which the store may be organized, for the purpose of this discussion we shall consider only two, namely, a continuous and a discrete store. These will be exemplified by the Rapid Selector and the Minicard System, respectively.

The Rapid Selector records the store on a continuous strip of film. Each item is entered on the film in sequence without any prefiling or organization of items. This means that a search involves scanning the total film.

The Minicard System, as its name indicates, makes use of discrete pieces of film which can be duplicated and prefiled in different locations in order to reduce the time of search. This means that the store is so organized that a search involves scanning only the appropriate section of the store. Search with the Minicard System is described by its constructors as follows:

> "To minimize the long search problem for the Minicard System, the file is set up on the basis of multiple entries, instead of a single entry for each document. Minicard duplicates - one duplicate for each significant code (term or word) contained in the Master Minicard - are made for each document.

Duplicates are sorted to separate sections of the file so that each file section consists of all cards containing a particular code (or in some cases, combinations of codes). For many questions, the requirements of a search are satisfied if only a single file section is presented to the selector. The search of a single file section can be accomplished in a few minutes." [2]

2. Multiple Searches and the Queuing Problem.

The problem of queuing in a machine storage and retrieval system is basic. It sets the requirement for time of search in terms of the number of searches to be made in a given period of time. This subject will be developed at length in other papers. Here it is only necessary to note that the different organizations of the store in the Rapid Selector System and in the Minicard System set up different possibilities for multiple searches. Since the whole file must be scanned in any single search with the Rapid Selector, multiple searches are possible only if the selector has sufficient electronics and memory to scan the store with multiple questions in parallel. In other words the form of the store in the Rapid Selector indicates that any requirement for multiple searches must be met by increasing the number of terms that can be searched simultaneously by the selector.

The Minicard System, however, can provide for multiple searches by increasing the number of individual selectors, since as indicated above any individual search involves only the search of a single file section. Thus many searchers could have access to the system at the same time if multiple selectors were available. The converse of these statements is also most important. Given the organization of the store in the Rapid Selector, multiplying the number of selectors would not be effective. And given the organization

of the Minicard store, increasing the electronics (number of potential terms in the question) in a single selector is likewise ineffective.

Determining the minimum, mean, and maximum size of a question which will be put to any storage and retrieval system is not a theoretical problem but a problem the answer to which is delivered by experience. But once experience has supplied this information, it is apparent that providing electronics in a selector for questions larger than any which will ever be asked is bad design. It is like building a 14-inch main for a one-inch flow of water. It can be asserted that most questions involve two or three terms; some may involve four; five term questions are very rare; and questions involving six* or more terms are so rare that they can be assumed not to exist. The only reason, then, for building a selector which can ask a question containing more than five terms would lie in the possibility of programming the selector for multiple searches during a single scan of the store. If this reasoning is applied to the Minicard System, it becomes apparent that the electronics in the present selector which can handle a 20 term question could provide for 5 selectors each capable of handling a 5 term question, that is to say, 4 terms in the selector plus the term which designates the file section to be searched. ("In as many cases as possible Minicards to be searched will be cards in a file section representing a code which is related by conjunction with the rest of the question."[3] This means that the Ministick filing term is also available as a term in the question, in addition to the terms in the selector.)

---

*An examination of the Decennial Index to Chemical Abstracts indicates that an index heading involving five or six terms is sufficient to select a unique item from over half a million items. Since there are only about 1,000 items on each Ministick, searching a Ministick by a question of 5 or 6 terms is a supererogatory search.

A single selector capable of handling 5 questions at once fulfills no purpose in the Minicard System because the organization of the store makes it useless to search a single file section for multiple questions. The prefiled store of the Minicard System does make it possible, as has been noted above, to use many selectors in parallel as contrasted with parallel questions on the same selector. It can be concluded therefore that replacing the present Minicard Selector with 5 smaller selectors having a total circuitry which is approximately the circuitry in the present selector, will increase the output efficiency of the system 5 times at no appreciable increase in cost. Indeed it would be quite reasonable to design the selectors so that 8 of these handled 3-term questions (2 terms in the selector and one in the file unit); and only one handled a 5-term question (4 terms in the selector and one in the file unit) making a total of 20 terms provided for by electronic circuitry. This would result in a nine-fold increase of efficiency of the system in selection or output.

It must not be supposed that the relation of the size of the question to the work accomplished and the amount of circuitry required applies only to the Minicard System. The conclusions presented above and the mathematical analysis in the Appendix apply quite generally to any device using direct free field coding.

References

1. Kreithen, A., AFOSR TN 57-400 or ASTIA AD 132475, June 1957.
2. Kuipers, J. W., Tyler, A. W., and Myers, W. L.: A Minicard System for Documentary Information, in "Symposium on Systems for Information Retrieval", Western Reserve University, Cleveland, Ohio, April 15-17, 1957, pp. 10-11.
3. Ibid. p. 15.

APPENDIX

I Definitions

A. The unit of retrieval or search work is one word of the store collated with one word of the search question.

B. The unit of search power is one search work unit per unit time.

II Sequential Search Work for Term Combinations

A. Assume a store or memory of N items indexed by an average of n terms each. The search is made over one item at a time, i.e., it is a search of n words one at a time, repeated N times. Another way to say this is that the search capacity is one word of a question consisting of Q words.

B. The work per item in search for the first word is $1 \cdot n$, and for N items it is nN.

Assume that a fraction p_1 of the N items is selected by the first word. That is, Np_1 items are selected and will be searched for the second word.

The work on the second word is

$$1 \cdot n \cdot Np_1$$

Assume a fraction p_2 is selected by the second word. The work on the third word is

$$1 \cdot nNp_1p_2$$

For the Qth word the work is

$$1 \cdot nNp_1p_2 \text{---} p_{Q-1}$$

C. The work in selecting Q words sequentially is

$$\begin{aligned} W_s &= Nn + Nnp_1 + Nnp_1p_2 + \text{---}\ Nnp_1 \text{---} p_{Q-1} \\ &= Nn\left[1 + p_1 + p_1p_2 + \text{---}\ p_1 \text{---} p_{Q-1}\right] \\ &= Nn\left[1 + \sum_{i=1}^{Q-1} \mathbb{P}\,(p_1p_2 \text{---} p_i)\right] \\ &= Nn\left[1 + \emptyset\right] \end{aligned} \tag{1}$$

III <u>Parallel Search for Term Combinations</u>

A. Assume a search capacity of Q words. The work per item is Qn, and is always the same. For N items it is

$$W_p = NnQ \ * \tag{2}$$

IV <u>Comparison of Parallel and Sequential Search Work</u>

A.

$$\frac{W_p}{W_s} = \frac{NnQ}{Nn[1+\emptyset]} = \frac{Q}{1+\emptyset} \tag{3}$$

Now $0 \leq p_i \leq 1$, and <u>at least one</u> $p_i < 1$. Therefore

---

* A memory of 2 bits is also required to register the success of finding A and finding B. But the relation of memory to units of work is omitted here.

$$1 + \sum_{i=1}^{Q-1} \mathbb{P}\,(p_1 p_2 \,\text{---}\, p_i) = 1 + \emptyset < Q$$

$$\frac{W_p}{W_s} > 1 \qquad (4)$$

More search work is performed by a parallel than a sequential search.

V Comparison of Parallel and Sequential Search Time

A. Let R_s be the sequential search work per unit time or sequential search power. Then $R_p = QR_s$ is the parallel search power.

The time of search is given by

$$T = \frac{\text{Work}}{\text{Search Power}}$$

B.

$$T_p = \frac{NnQ}{R_sQ} = \frac{Nn}{R_s} \qquad (5)$$

$$T_s = \frac{Nn(1+\emptyset)}{R_s} \qquad (6)$$

$$\frac{T_p}{T_s} = \frac{1}{1+\emptyset} > \frac{1}{Q} \qquad (7)$$

C. The overall work and overall time compare as follows:

$$W_p > W_s \qquad \text{(from 4)} \tag{8}$$

$$T_p > T_s/Q \qquad \text{(from 7)} \tag{9}$$

$$R_p = QR_s \tag{10}$$

The last three equations show that when the parallel search rate is Q times the sequential search rate,

(1) the parallel search work is greater than the sequential search work

(2) the parallel search time saved is less than the increase in search power required to save the time.

VI Search Work for Ordered and Non-ordered Question and Store

The effect of searching for a question in which the order of terms is specific, i.e., AB, but not BA.

To distinguish AB from BA it is necessary either (1) to increase the size of the term, that is, make the two words A and B the single word AB or (2) add counters to the internal memory to indicate the temporal order in which A and B have been matched.

(1) Increasing the size of the word

A word is defined as a set of characters in a fixed order. Hence the requirement that words in a question be searched in a fixed order converts the words into one large word. The terms in the item must then be scanned in combinations as large as the number of words in the question.

There are $n!/Q!(n-Q)!$ such combinations per item. The units of work per item are

$$Q \frac{n!}{Q!\,(n-Q)!} = Q \binom{n}{Q}$$

For N stored items the total work for an ordered question is

$$QN \binom{n}{Q}$$

Since $QN \binom{n}{Q} \geq QN$, asking a question which involves a fixed order of terms multiplies the work accomplished by a factor which can be large.

(2) If a counter is added to the internal memory an ordered question can still be asked in QN units of work.

Conclusions:

(1) In any search of an information store by either sequential or parallel matching, the sequential search work is less.

(2) Search time saved by a parallel instead of a sequential search is proportionately less than the increase in search power required to save the time.

(3) Asking an ordered question involves either more units of work or more memory capacity than asking a question in which order of terms is immaterial.

CHAPTER IV

AN EVALUATION OF "USE STUDIES" OF SCIENTIFIC INFORMATION*

By Mortimer Taube

This paper attempts an evaluation of the total existing literature of use studies. It accepts the conclusions drawn by other surveys of use studies which appeared before the International Conference on Scientific Information in November, 1958, and brings these conclusions up to date by abstracting and evaluating the studies prepared for the Conference.

An attempt is made to analyze the reasons for the generally accepted failure of use studies by establishing a distinction between consumer services and professional services. It is concluded that the organization and dissemination of scientific information is a professional activity, the value of which cannot be measured by consumer responses, and that such responses cannot supply directions for the design of more effective scientific information and reference systems.

---

* Reprinted from the report submitted under the same title to the Information Complexes Division, Directorate of Mathematical Sciences, Air Force Office of Scientific Research, Contract No. AF 49(638)-91, December 1958.

The papers prepared for the International Conference on Scientific Information are divided into seven groups, each group covering a specific defined area. Area 1 is concerned with "literature and reference needs of scientists; knowledge now available (of such needs), and methods of ascertaining requirements."[1] In the guide to the scope of this Conference Area, it is stated that, "In order to improve the dissemination of scientific information and to design more effective reference tools and services, we need to have a more complete understanding of the weaknesses and strengths of the present pattern of scientific communication and, in particular, of the unfilled or inadequately filled needs of scientists for information."[2] In the evaluation of papers prepared for this area and similar papers, which shall henceforth be referred to as "use studies," it is important to keep this well-defined purpose in mind.

Coming as these papers do at the beginning of the Conference program, it is clear that the planners of the Conference hoped that the requirements set forth in this area would supply criteria in terms of which actual or proposed systems of disseminating and storing scientific information described in subsequent areas could be measured. It is clear that this purpose, stated so clearly, supplies a criterion against which not only the use studies prepared for the Conference but all similar studies can be evaluated.

It is important to note that this criterion for the evaluation of use studies, namely, the extent to which they help "to improve the dissemination of scientific information and to design more effective reference tools and services" is not the only possible basis of justifying and evaluating use studies. Use studies might be carried out as pure descriptions of scientists' behavior without any other motivation; or they might be designed to lead to an improvement in the attitude and behavior of the scientist with reference to information which is available to him. This point is not facetious; the

studies of the use of library catalogs by investigators within the library profession, from which use studies of scientific information are a natural descendant, had dual purposes. On one hand, it was hoped that such studies would provide information concerning methods for improving catalogs; and on the other hand, it was hoped that they would provide information concerning the need for improving (i. e. training) users. However valid this latter purpose, it will not be considered further as a criterion to evaluate the use studies prepared for the Conference and their prototypes.

As a background to her paper,[3] prepared for Area 1 of the Conference, Tornudd has listed sixty-nine previous studies. Of these sixty-nine studies, only fifteen, which date from 1955, are not listed in previous compilation by Shaw,[4] Egan and Henkle,[5] and Stevens.[6] All major use studies were analyzed and abstracted by Shaw as part of his work for the National Science Foundation. Hence, Shaw's evaluation[7] can be considered to be based on all data available at the time of the preparation of his report.

It has been recognized above that studies of the use of library catalogs have a close relationship to studies of the use of scientific information; and a summary treatment and evaluation of studies of catalog use, prepared by Frarey,[8] is available. Thus, Frarey and Shaw provide an evaluation of all the data available up to the papers prepared for the Conference or listed by Tornudd as having appeared after the completion of Shaw's compilation.

What these bibliographical facts portend is that studies of catalog use and use studies of scientific literature are of recent origin; and it is possible to cover the whole existing literature in a summary evaluation.

Frarey identified and described twenty-seven studies. He recognized the possibility that some studies may have

been "unwittingly overlooked," but he concluded that it is "safe to say that these twenty-seven contain the substance of what is presently known."[9] These studies give us information on difficulties which occur in using catalogs, which are traceable both to deficiencies in the catalog and deficiencies in the users; but they supply no basis for any serious modification of cataloging practice. Indeed, Frarey concludes that no amount of similar studies based upon quantitative measurements of users' habits could ever supply information upon which extensive modifications of the catalog could be based.[10] Thus, on the basis of the specific criteria established by the Conference, Frarey concludes that the studies surveyed by him have zero value and that similar studies have an expectation of zero value. Frarey proposes that qualitative studies of catalog use are necessary as a basis for conclusions concerning desirable modifications of the catalog.[11] But he is not clear concerning the nature of such studies. It is interesting, therefore, that Lilley, on the basis of a careful review of Frarey's work, and his own independent observations, concluded "that such studies (i.e. qualitative studies) are not only impossible, but would serve no useful purpose if they could be accomplished."[12] No studies have appeared in the literature which challenge Frarey's and Lilley's conclusions. Hence, both the results and the prospects of quantitative or qualitative studies of cataloging use seem to be of no value measured against the criteria of prescriptive value; that is, they have not led and cannot lead to any conclusions concerning desirable modifications of cataloging principles.

If, at this point, attention is turned from studies of catalog use to use studies of scientific information, the same negative conclusion emerges. Thus, Shaw, on the basis of a thorough analysis of all existing use studies states that, "None of the reports made to date provides a firm basis for planning (i.e. improving) communications programs."[13] Shaw believes the difficulties encountered by use studies are largely

methodological and that the specific data reported by the several studies are not reliable. He does not attack the more basic problem with which this paper is concerned: even if use studies could produce reliable data, can such data supply a valid basis for improving scientific information systems? This question, it will be recognized, is not concerned with the merits of any particular study per se, but with the validity, in principle, of use studies as guides to the improvement of information services. It was clearly on the assumption of such validity that the Conference planners established Area 1.

There are appended to this paper abstracts of the thirteen papers in this area submitted and accepted by the Conference. In effect, this appendix brings Shaw's work up to date and supplies a definitive body of data to be evaluated. Within the body of this paper, reference will be made only to particular points included in the Conference papers which have a direct bearing on the issue.

Tornudd concludes that "none of the methods used in studies on the use of information by scientists has proved to be truly reliable."[14] However, she expresses a hope that "the results of the operational research program underway in the USA should reveal better methods for the study of these problems."[15] Again, it must be remarked that increasing the reliability of data does not ipso facto increase their significance. This point is brought out clearly by the preliminary report on the Operations Research Study to which Tornudd refers, which was submitted to the Conference.[16]

The following rationale of the operations research program is presented: "The question originally asked of the Operations Research Group at Case by the Office of Scientific Information was: What is the possibility of applying Operations Research to problems in the dissemination of recorded information? The research reported here is a partial answer to this

question. To understand its development, it is helpful to be aware of two essential characteristics of Operations Research. The first is that Operations Research is concerned with the application of scientific method to the study of systems of organized activity rather than to the components of such activity. Its orientation is whole-istic. Secondly, Operations Research is operationally oriented. This means that its primary concern is with affecting the way systems operate and not merely in providing interesting information. In brief, it seeks to provide a basis for effective action."[17]

In other words, the object of the study is not to gather "interesting information" or "reliable information" but information which provides a basis for action. To determine how the effectiveness of a scientific information system could be increased, "a measure of effectiveness of the system is required."[18] It is assumed that "the system is concerned with increasing scientific productivity"[19] and hence that an increase in effectiveness of the system can be measured by or be considered equivalent to an increase in scientific productivity. Having arrived at this conclusion, the research team reasoned further that "an acceptable measure of scientific productivity was not likely to be obtained within the time available for the project."[20] This impasse is avoided by finding some other measurable characteristic which can be taken as an index of scientific productivity. "Since we could not expect to measure scientific productivity directly, we sought an aspect of scientific activity which (1) would be measured objectively and (2) if increased, would also increase scientific productivity. The time available for scientific research is such an aspect of scientific activity."[21] This conclusion led to a study of, "(1) How do scientists actually spend their time? (2) In what types of scientific activity are there the greatest potentialities for reducing time expended without reducing scientific output? How can these reductions be realized in the most effective way?"[22]

There is a certain plausibility in this conclusion, because it is a recognized aim of searching systems to reduce the time of search. Presumably such a reduction would make more time available for scientific productivity. But this plausibility vanishes when it is realized that the total elimination of searching for references or reading them will make more time available for scientific productivity than any mere increase in the efficiency of information apparati. If it is desired to maximize the time available for scientific research, then it also becomes desirable to abolish all journals and information systems, since consulting these must cut down the time available for scientific research.

It is apparent, at this point, that the authors have overshot the mark. They have identified scientific productivity with productive science. A group of research workers busily and ignorantly duplicating one another's work and writing articles in journals which nobody reads are contributing to scientific productivity but they are not, thereby, productive scientists.

What this means is that measures of effectiveness must involve judgments of value. For example, the time available for good, productive, scientific work might be increased by cutting down time wasted with inefficient reference systems; but it could not be increased by abolishing scientific periodicals and scientific communications. It is assumed that the resultant time spent in bad, duplicative, unproductive, scientific research would reduce good scientific production more than the abolition of reading would gain for it. To the extent, then, that this operation research study concerns itself only with increasing research time rather than increasing useful or productive research time, it cannot supply any measure of the effectiveness of a scientific communication system and has supplied no reason for modifying the negative conclusions of Frarey, Shaw,

Lilley, and Tornudd.

Many of the studies in Area 1 (Menzel,[23] Herner,[24, 25] Scott, Scott,[26] Glass and Norwood[27]) emphasize the unplanned nature of scientific communication. Data gathered by diary, interview and questionnaire supported the fact that "conversations with fellow scientists" and "accidental reading" are among the major sources of scientific information. The probability of "unplanned communication" can be increased by providing favorable circumstances for it to occur; but such conclusions cannot lead directly to improved services. There is, however, an oblique sense in which the prevalence of unplanned communication sheds light on the problem of making scientific communication more effective. Bernal, who on other occasions has suggested improvements in the present, almost random organization of primary publication, regards the prevalence of "unplanned communications" and "hit-or-miss reading" as evidence that primary dissemination of scientific information has broken down and that "it is in present conditions growing more and more difficult, and may soon be impossible to disseminate scientific information unless that is done in a fashion that permits its easy storage or at least its easy processing through a storage and retrieval mechanism."[28] When the organization of primary distribution reaches a state of complete entropy with reference to the reading habits of scientists, that is, when unplanned communication becomes as effective as planned communication then dissemination in any meaningful sense will cease. Storage and retrieval systems will be required to restore organization to the primary publications in order to make planned reading and study possible.

As a final measure of the difficulty of regarding use studies as a guide to the improvement of information services, Bernal's comment on a finding of Urquhart's is very revealing: "Dr. Urquhart has shown that out of 9100

periodicals taken by the Science Library in London, 4300 were not consulted at all in a given year. Now it is difficult to believe that nothing of interest to the 87,000 readers at the Science Library was to be found in these 4300 periodicals. If so, the sooner they cease publication, the better."[29]

The one immediate conclusion which would seem to emerge from use studies, if we take them at their face value, is that scientists don't read and that scientific writing and publication is a largely redundant enterprise. If the needs of scientists for scientific information are derivable from the use they make of it as revealed in these studies, then these needs are very minor indeed, and the indicated course would be a virtual moratorium on scientific publication, abstracting, indexing, and the like.

An interesting question remains: Why do use studies continue to accumulate in the face of a general recognition that they have so far contributed little to the purpose for which they are primarily intended? The answer to this question is to be found in the fact that use studies have been criticized in terms of their methodology (diary, questionnaire, case study, interview) reliability, thoroughness, special character, etc. With the exception of Lilley, the principle and rationale of use studies has never been directly examined. So much of social activity seems amenable to polls, surveys, questionnaires, etc., that it seems reasonable to apply similar techniques to the problems of scientific information. Why, then, have use studies uniformly failed to provide a measure of effectiveness of scientific information systems?

The answer to this question can be found in the distinction between consumer services and professional services. Consumer services can and should be evaluated in terms of consumer response. Any organization offering such a service

must study and maintain an awareness of consumer use, indifference, or rejection of its services. On the contrary, a professional service differs from a consumer service in possessing criteria of evaluation which are independent of consumer response. A professional service may engage in various activities to acquaint consumers with the value of its services, but this value is not measured by consumer acceptance.

The sale of packaged breakfast foods is a typical consumer service. If consumers reject or are indifferent to a particular product, the product is worthless. One breakfast food can be substituted for another, and it would be just silly for a company to insist that consumers buy and eat a breakfast food they did not like.

Medicine is a typical example of a professional service, and the recent experience in this country with the Salk vaccine presents an illuminating and illustrating fact that the value of a professional service cannot be measured in terms of consumer response. Certainly, it was necessary to persuade consumers to use the vaccine; and private, state, and federal agencies engaged in protracted campaigns to persuade consumers that the vaccine would help them. But no one proposed that the value of the vaccine be measured by a use study; that is, by asking consumers whether they liked or didn't like the vaccine.

There is, of course, an important sense in which the user plays a role in the evaluation of the Salk vaccine or any other medical service. The value of the vaccine is measured by the reduction of polio among users of the vaccine compared with the incidence of polio in a control group to whom the vaccine is not given. The measure of value here is the incidence of polio. If there existed a standard of good scientific production, then it might be possible to measure

the value of a scientific information service by providing it to one group and withholding it from another control group. It is interesting to note that Bernal has proposed such a study. "In addition, I have proposed a competition in scientific research in the same field of three teams (a) with the best available information services, (b) with present average information services, (c) with no information services at all."[30]

It is the contention of this paper that the provision of scientific information services is a professional activity and hence the value of such services cannot be measured by use studies. Those groups and individuals who have sponsored and carried out use studies are denying this professional character of scientific information service. They are, in effect, equating the work of editors, indexers, abstractors, authors, systems designers, etc., to that of purveyors of packaged tidbits whose value is measurable in terms of consumer response to their delectability.

To be sure it is more difficult to measure the value of a scientific publication (that is, the value of reading it) than it is to measure the value of health as compared with the value of polio. But, unless we are prepared to say that a scientific publication <u>should</u> be read, that a scientific information service <u>should</u> be used, the very enterprise of scientific communication ceases to be significant. Certainly, some scientific information services are good and some are bad; but the measure of such value is found in the professional competence of the scientist as editor or author and of the information specialist. It cannot be found by counting noses of users and non-users or by discovering that scientists being normally lazy would rather ask their colleagues for information than to take the trouble to dig it out of the literature.

In one of the studies prepared for another area of the

Conference, a justification for use studies is based upon a categorical denial that any professional competence or standards exist which could be used to evaluate an information retrieval system. "At present, absolutely nothing can be taken for granted; there is no single fact which can be demonstrably shown to be true; no theory put forth by one expert which is not refuted by another."[31]

Certainly, if such a statement is taken seriously, the holding of the Conference itself becomes of dubious value or rather it becomes a meeting not of experts, but of salesmen. As a matter of fact, the statement is the typical nonsense used to justify a vaguely conceived use study. As an example of "one single fact which can be shown to be true" consider the following: It is easier to find a particular name in an alphabetical array of names than in a random arrangement. This is certainly a fact so trivial or rather so generally accepted that it is hardly ever mentioned. It becomes relevant only because of such statements as "absolutely nothing can be taken for granted," not even, one may suppose, the known order of the alphabet.

There is one final point which must be considered in evaluating use studies. Sometimes such studies are concerned with the use of primary publications; that is, journals, texts, data compilations, reports, etc. Sometimes, they are concerned with what have been called secondary systems; namely, indexes, abstracts, reviews, bibliographies, etc. And sometimes, they have been concerned with both aspects of the total scientific communication system.

Actually, a sharp distinction must be made between primary and secondary materials in that primary materials are presumed to have intrinsic value whereas secondary materials have only an instrumental or means value. The value of an index to a journal derives from the value of the articles in the journal. If no one has any interest in the

articles, then certainly an index to them is supererogatory. The use studies which were prepared for and discussed at the Royal Society Scientific Information Conference in 1948 were largely concerned with primary publications. The Bernal scheme to modify primary publications became the "cause célèbre" of that conference; and there was no emphasis upon use studies as supplying a guide to the design of reference, abstracting, cataloging, indexing, and bibliographical systems.

The emphasis on use studies as a basis for the design of more effective secondary systems of scientific information is a recent phenomenon. It is perhaps unfortunate that the Conference papers which constitute the most recent attempt to determine the value of information systems on the basis of use studies actually resulted in an attack upon the whole enterprise of primary publication. These studies taken at face value disclosed that indexes, abstracts, bibliographies, and reference services weren't used; not because they lacked instrumental value, but because scientific publications lacked intrinsic value. No use study submitted to the conference disclosed that poor abstracts or indexes were a barrier to getting primary materials that were wanted; rather they seemed to indicate that nobody was very much interested in the primary publications that were so laboriously indexed and abstracted. If any conclusion were to be drawn from these studies, it is not that we should stop indexing, but that we should stop publishing. Surely, this was not the intent of the planners of the Conference.

The inevitable conclusion of this paper is that use studies have no value as direct guides to the design of information systems, any more than consumer acceptance or rejection is a guide to the value of the Salk vaccine. But this does not mean that use studies may not have other forms of value. A study of the behavior of scientists may help provide information concerning the necessity of sugar coat-

ings on knowledge pills. Consumer resistance to swallowing capsules over a certain size may give valuable clues to the optimum size of information packages. Further, such studies, as in the case of the Salk vaccine, may indicate that consumer education or training is necessary if the full value of scientific information systems is to be made available to society.

The design of such systems remains a matter of professional competence. All phases of such systems should be studied from the writing of scientific papers through their publication, dissemination, storage and retrieval to their use. The documentalist or information specialist is essentially an engineer. He designs a system in which a favorable ratio is achieved between input costs and pay-off in potential use. Part of his design may call for a program to train or retrain users so that they may secure the maximum information potentially available from the system. He will certainly be concerned with patterns of initial dissemination of primary material in order to determine whether a storage and retrieval system should be planned for a situation of organized or disorganized dissemination. Bernal noted that disorganized primary dissemination means that a storage and retrieval mechanism must be created between the publisher and the user. But this relationship implies that a better organized system of primary dissemination will reduce in some measure the demands made upon storage and retrieval systems. Whether the area of maximum pay-off is to be found in organizing primary publication or in planning storage and retrieval systems to handle random initial dissemination is a topic which is now being investigated and which will be the subject of subsequent papers.

References

1. International Conference on Scientific Information, Washington, D. C. November 16-21, 1958, Preprints of Papers. Washington, National Academy of Sciences, National Research Council, 1958. Area 1, p. 3; hereafter referred to as Conference Preprints, Area 1.

2. Ibid. p. 4.

3. Tornudd, Elin: "Study on the Use of Scientific Literature and Research Services by Scandinavian Scientists and Engineers Engaged in Research and Development." Conference Preprints, Area 1; pp. 9-65.

4. Shaw, R. R.: "Studies on the Use of Literature in Science and Technology." In Pilot Study on the Use of Scientific Literature by Scientists. Washington, National Science Foundation, 1956.

5. Egan, M. and Henkle, H. H.: "Ways and Means in Which Research Workers, Executives, and Others Use Information." In Documentation in Action. N.Y. Reinhold, pp. 137-159, 1956.

6. Stevens, R. E.: "Characteristics of Subject Literatures." ACRL Monographs, Nos. 5-7, pp. 10-21, January, 1953.

7. Shaw: Op. Cit. p. 1.

8. Frarey, C. J.: "Studies of Use of the Subject Catalog: Summary and Evaluation." In The Subject Analysis of Library Materials, edited by Maurice F. Tauber. New York Columbia University School of Library Service, pp. 147-166, 1953.

9. Ibid. p. 150.

10. Ibid. p. 154.

11. Ibid.

12. Lilley, O. L.: "Evaluation of the Subject Catalog: Criticisms and A Proposal." American Documentation Vol. 5, No. 2, April, 1954, p. 44.

13. Shaw: Op. Cit. p. 2.

14. Tornudd: Op. Cit. p. 62.

15. Ibid.

16. Halbert, M. H. and Ackoff, R. L.: "An Operations Research Study of The Dissemination of Scientific Information." Conference Preprints, Area 1, pp. 87-120.

17. Ibid. p. 87.

18. Ibid. p. 88.

19. Ibid.

20. Ibid. p. 89.

21. Ibid.

22. Ibid. pp. 89-90.

23-27. Cited in appendix.

28. Bernal, J. D.: "The Transmission of Scientific Information: A. User's Analysis." Conference Preprints, Area 1, p. 77.

29. Ibid. p. 70.

30. Ibid. p. 85.

31. Cleverdon, C.: "The Evaluation of Systems Used in Information Retrieval." Conference Preprints, Area 4, p. 35.

APPENDIX

Abstracts of Preprints of Papers for Area 1
International Conference on Scientific Information
National Academy of Sciences - National Research Council
Washington, D. C. - 1958

1. Tornudd, Elin: Study on the Use of Scientific Literature and Reference Services of Scandinavian Scientists and Engineers Engaged in Research and Development. pp. 9-66

This study contains a bibliography of previous studies based on the compilations of Shaw, Egan and Henkle, and Stevens; the study proper was based upon 200 questionnaires filled in by 100 young Danish scientists and 100 young Finnish scientists. Based on completed questionnaires, tables were compiled showing: (1) Distribution of respondents by field of research and institutional affiliation, (2) Estimated ability to keep up with new developments, (3) Source of information, (4) Time devoted to literature research as related to various factors, (5) Subscription to and use of journals, (6) Use of foreign languages, (7) Library availability and use, (8) Services needed, (9) Papers published and publishing media, (10) Difficulties in obtaining information, (11) Instances of duplication of research, (12) Suggestions for improving reference services and / or skill in using them. This paper concludes that even though all studies of use are unreliable, it is necessary "for every information service to carry out a continuing analysis of the requirements of its users, especially of its least industrious users."

2. Bernal, J.D. Transmission of Scientific Information pp. 67-86

This paper is not a use study, but a critique of their rationale:

"My main reason for presenting this paper at such a Conference is that I believe that the whole subject of transmission of scientific information needs an analysis of a descriptive or natural historical kind before we can hope to find the right figures to look for or the right questions to ask. This is not only to ensure that the answers we get are significant, statistically or otherwise, but also to determine whether the answers that prove to be significant and true are really relevant to the total situation that we hope to understand and control; namely, an improved flow of scientific information.

"The main reason behind this implied criticism is that if the matter be treated as one of operational research, it follows that all enquiries as to present uses of scientific information services, through a necessary background, can by themselves tell us little of use for improving the services. They tell us what people do with an admittedly very imperfect service, not what they would do with a better one (which would naturally include proper training for its users.) A certain amount could be learned by a comparison between different systems in use, and some lessons from this quarter may emerge from our Conference, but we cannot hope to learn much until it becomes possible to carry out trials involving considerable variations under strictly comparable conditions.

"The essential difficulty is that, though the user may well know what he wants from an information service, he is in no position to know what he needs from it, namely, what variation in the system would help most to further his work. Consequently, any action based on analysis of present user habits is unlikely to produce impressive results."

There is appended to the paper a list of suggested studies

which might be carried subsequent to a descriptive study of the transmission of scientific information.

3. Halbert, M. H. and Ackoff, R. L: <u>An Operations Research Study of the Dissemination of Scientific Information.</u> pp. 87-121

A preliminary study based on observation of activity of how approximately 1500 scientists spend their professional time. The data are reported for approximately 18,000 observations. This preliminary study is part of a larger study intended to provide data which can be used in making disseminating systems more effective.

4. Hogg, I. H and Smith, J. R.: <u>Information and Literature Use In a Research and Development Organization.</u> pp. 121-152

A uniform sample drawn from three arbitrary status grades of applied scientists (Research Managers, Senior Staff, and Junior Staff), totalling 157 persons, were interviewed using standard questionnaires, and also were given 14-day reading diaries to complete. Chief information sought was: (a) how they obtained their scientific and technological information and how they valued the different sources; (b) how many abstracts, periodicals, research reports, and textbooks they read during 14 consecutive days, and where they read them; (c) whether they considered they had adequate time for reading at work; (d) where they obtained the literature read during the 14 days, how they got references to it, and how much literature they bought themselves; (e) their criticisms of the various library lists as reference sources; (f) how they used the information gained during their 14 days of reading, and what reference-sources led to the most useful reading; (g) the value they placed upon periodicals according to age and language, and their use of those British and foreign origin; (h) whether they kept personal data records; (i) their

suggestions for new, and criticisms of existing library services; together with the formal qualifications and field of research of those interviewed. All the sample were interviewed, and 92% of the diaries were returned.

The first analysis of results showed that: (a) The prime information sources were informal contacts and the literature, of which reports were valued most and periodicals least by those answering. (b) Less than one abstract-consultation was made per head during the 14 days' reading; less than one-third of the sample read any abstracts during this period. Two-thirds of the consultations were for keeping up with the literature, one-third for locating past literature. All except one of the diarists read some periodicals, reports and textbooks, an average of 4 per head. Three-quarters of the reading was in working-hours, the rest at home. These figures are, however, statistically suspect. (c) Three-quarters of those interviewed said that for some part of the past working year, they had no time to read in working hours [cf. (b)]. (d) Over half of the diarists' literature was from the Group libraries, one-quarter, mainly reports, was sent direct by authors or colleagues, one-seventh was the diarists' property, and a small amount was borrowed from colleagues. Less than half the sample bought their own books, one-third bought their own journals; average spending per head in the past year was Ł 4 on books, Ł 2 on journals. Of the 14 days' reading, no references were required for 40% of it; of the remainder, colleagues recommended 18%, the diarists' memory or knowledge accounted for 18%, the library provided references for 14%, references in other publications were 6%, and abstract journals (and the library catalogue for books) provided 4%. (e) Of those interviewed, 25% criticized the library bulletin (of selected journal references), 13% book accessions list, and 17% the report list. All lists were used by about three-quarters of the sample. The main tendency was to ask for more selective lists for individuals or their immediate departments, and

for more abstracts or annotations in the lists. (f) Half the diarists' reading was in aid of this research work, one-third for general interest, and only about 3% was discarded as of no use. The highest percentage of "discarded" per reference source occurred when the least-used source, abstract journals, was used (about 14%.) (g) With a maximum usefulness-score of 6, among those who used them current periodicals of research scored 4.6, falling after 10 years to 2.9; current periodicals of technology scored about the same. Decline of interest in both types of periodicals was heaviest among engineers, least among chemists and metallurgists. Of foreign language periodicals, among those who used them the "face value" (including the effect of the language-bar) was lowest for the Japanese periodicals; the potential value (assuming no language-bar) was the highest for the German; and the difference between face-value and potential value was highest for the Russian. Of periodicals of British and foreign origin, 57% of the scientists' information came from the former, and 42% from the latter. (h) Personal records of data or useful references were kept by two-thirds of the sample, and another one-tenth of them used records kept by others in their section. (i) Two-thirds of the sample offered criticisms of or suggestions for improvement of the library service; only those of wider interest are mentioned. Major comments were: (a) the libraries should publicize their services, (b) more or better qualified library staff are needed, (c) librarians should notify their users individually of literature of interest, (d) better copying (reprint) facilities should be provided. [Author abstract]

5. Fishenden, R. M.: Methods by Which Research Workers Find Information. pp. 153-170.

A survey has been made at the Atomic Energy Research Establishment, Harwell, to discover the methods by which research workers obtain the information they use and read.

The object of the survey was to find which methods were most effective in bringing information to their notice, and so to improve the information services in the establishment. The survey was made by two methods; diary cards and personal interview.

The results showed that the following were the principal ways by which information was found: The figures represent percentages of all items recorded in the diary survey, regular reading of the current literature, including new reports, 29%; papers found through references in other papers, 9%; personal recommendation, 11%; and scanning lists of titles included in the report lists and information bulletin issued by the library, 17%; Nuclear Science Abstracts, 7%; found for readers by the library, 4%.

For the retrieval of old information (22% of all items recorded), there was a marked reliance on personal indexes (4%) and "previous use" (i.e. Memory) (10%). All other retrieval methods combined accounted for only 8% of items. There is a strong inference that inadequate attention is paid to systematic searching of the literature and that greater use could be made of library services for such services.

The use of the foreign language literature was small (5%) as was the use of reviews (4%).

Comparison with other records, comparison with the diary and interview surveys, and the general consistency of the figures indicates that the results of the diary survey were unexpectedly reliable. An important conclusion is that useful results can be obtained from a much simpler diary card than those used in some previous investigations.

The detailed results, relating as they do to a particular set of circumstances, are of limited general interest, but they give valuable and much needed guidance on the ways in

which the AERE information services should be developed. [Author abstract]

6. Herner, S. and Herner, M.: Determining Requirements for Atomic Energy Information from Reference Questions pp. 171-178.

The method of determining information requirements for reference questions, "has its pitfalls and limitations"; but the paper gives no indication concerning a method or technique for overcoming such difficulties. Since most of the reference questions studies involved only "logical products" of two or three terms, it is concluded that these questions could be answered by a system which provided answers in terms of logical products of two or three terms. The major conclusion of the paper is that the study resulted in useful data.

7. Spirit, J. and Kofnover, L.: Systematically Ascertaining Requirements of Scientists for Information pp. 179-184.

This paper describes a method of indexing the interests of research workers by UDC numbers. Material received by the information center is analyzed in accordance with the UDC system. This analysis then provides for automatic internal dissemination to research workers since the same class numbers relate their interests to relevant incoming material.

8. Glass, B. and Norwood, S.H.: How Scientists Actually Learn of Work Important to Them pp. 185-188.

A pilot study based on interviews with 50 scientists to ascertain how they learned of work crucial to their own. A table is presented ranking the methods. Casual conversations

rank first.

9. Menzel, Herbert: Planned and Unplanned Scientific Communication pp. 189-234.

Under a grant from the National Science Foundation, Washington, D. C., the Bureau of Applied Social Research of Columbia University has undertaken to explore ways in which communication research by interview survey methods can contribute to an understanding of the needs and means of scientific information-exchange. On the basis of such an understanding, proposals to improve scientific communication might be generated and evaluated. As a first step, it was decided to study the information-exchanging behavior of the biochemists, chemists, and zoologists on the faculty of a single academic institution -- a prominent American University. This paper reports selected results. A more complete account is on deposit with the National Science Foundation.

The exploratory study was intended to define problems, categories, and procedures for more systematic investigation. Although this report contains numerous frequency counts based on interview responses, they are to be regarded as illustrations of the possible outcome of further work and not as reliable findings. They may not even reliably describe the three academic departments studied, since the interview schedule was continuously modified and developed as the work proceeded. [Author abstract]

10. Scott, Christopher: The Use of Technical Literature by Industrial Technologists pp. 235-256.

The basic assumption of this paper is that "we had better take the scientist as we find him and build our systems of information storage around him." Interview techniques were

used with 1,082 industrial technologists of which only 17% had any degrees and 61% had no academic or technical qualifications. The survey disclosed very little reading and most information gained from reading was gained accidentally; that is, it was not being deliberately sought by the reader. The reference use of literature is very much less significant. Acknowledgement is made in the text to Mr. Herner for some of the questions used in the interview.

11. Spurr, S. H.: Requirements of Forest Scientists for Literature and Reference Services pp. 257-266.

An examination of the need for information services and the kinds of literature which should be, but is not used by forest scientists. The author recommends a method of establishing a card file of forestry literature based upon the Oxford Forestry Classification. The author notes a serious problem of determining card size; since 4x6 cards are probably too small, 5x7 cards are recommended.

12. Herner, Saul: The Information-gathering Habits of American Medical Scientists pp. 267-276.

Five hundred scientists were interviewed concerning their "Information gathering habits." The author recognized that questions have been raised concerning the validity of face-to-face interviews. He notes that the Survey Research Center of the University of Michigan approves of survey interview techniques.

The primary conclusion of the survey "is a reaffirmation of the significant role of personal contacts in getting and transmitting scientific information." Informal tools are used for informal searches; formal tools are used for formal searches.

CHAPTER V

THE COMAC: AN EFFICIENT PUNCHED CARD COLLATING SYSTEM FOR THE STORAGE AND RETRIEVAL OF INFORMATION*

By Mortimer Taube

[This paper and the following paper by Mr. Murphy should be read in conjunction, since the Murphy paper describes the first reduction to "hardware" of the COMAC idea. In bringing about this reduction to practice, the IBM Company abandoned the type of coding originally suggested in this paper in order to secure compatibility with the family of IBM machines.

This reduction of efficiency in coding means that the estimates of searching times and reduction of storage space for the index given in this paper were not realized in the device described in the next paper. This statement is not intended as a criticism of the IBM Company. The speed with which the IBM Company produced the Special Index Analyzer is truly amazing and the company is certainly to be congratulated for its accomplishment in this regard.]

The concept of item codes and term codes has been fully

---

* Reprinted from the report submitted under the same title to the Information Complexes Division, Directorate of Mathematical Sciences, Air Force Office of Scientific Research, Contract No. AF 49(638)-91, October 1957.

developed in other papers; and from these papers we take the conclusion that there are only two basic patterns of grouping codes in a store: Either term codes are collected under item codes or item codes are collected under term codes . We also utilize the conclusion that, in a system of direct free field coding, a search consists of matching (or collating) a code (or codes) in the question successively against codes in the store.

Although these theoretical conclusions are generally applicable to any type of storage and retrieval device, in this paper their implication will be applied to the design of a specific system, namely, a new system of punched card collation which we have designated the Continuous Multiple Access Collator (Comac).

Since it is the purpose of this paper to demonstrate that the Comac is an efficient device for even the largest collections of information, e.g., patents, intelligence files, newspaper morgues, and picture files, we will base our calculations of search time and size of the store on the following figures:

Number of items in the store, 1,000,000.

Average number of term codes used to index an item, 20.

Number of terms in the vocabulary or different term codes used in the system, 10,000.

Within the general concept of matching we distinguish two methods employed by standard punched card devices. These two methods are usually shown as (1)searching and (2)collating.

Searching

Searching is performed with a sorter by making several successive sorts until all items coded by a certain term or

terms have been selected, that is, sorted out from the rest of the deck. Changing the column selector in the sorter and selecting the cards from the proper pocket constitute setting up a question in the reading head of a piece of apparatus, and this question is matched successively against codes in the store. It is assumed that each item is represented by a card or set of cards on which are grouped the term codes characterizing that item.

With standard punched card equipment (unless superimposed punching or wiring is used) the term codes in the question must be matched against specified fields on the item cards. For example, a simple sorter "sorts" cards column by column as determined by setting the column selector in the sorter; and the "101" machine which can search many columns at once must be programmed to search in the proper columns for term codes in the question (the "101" can be wired to search for a single code in any of a number of fields).

The IBM Corporation has constructed an experimental searching device which can search for multiple codes any place on a card. This device, which uses direct free field coding, is known as the Luhn scanner; it represents a significant advance in punched card searching.

The operation of the Luhn scanner is illustrated* in Fig. 1.

Each item card in the store is coded with the terms characterizing the items and the unused columns of the card are

---

* This illustration is only theoretically accurate. Since the scanner, as constructed, reads only one-half of the card, lacing is required on only one-half of the card area. Also, by virtue of always using codes of 5 out of 12 holes, lacing can be accomplished by just one extra punch which will always let light through.

"laced", that is, all holes are punched. The question card is also laced except for the columns required for the term codes in the question. The coding of a term in the question is the complement of the code of a term in the store (Fig. 2).

It is apparent that as a card in the store passes over the question card, the matching of complementary codes will cause a "blackout" which can signal a photocell to actuate a selection mechanism. The use of complementary coding and "blackout" cuts down the requirement for reading apparatus to one photocell per column per code area. If direct matching of codes were employed, each hole on the card would have to be read for a match or failure to match. (It will be seen that the use of complementary coding is only possible when a question card is prepared for a specific search and cannot be used in collation in which any card in the store may be used as a question card.)

We will assume here that each card in the store to be searched by a Luhn scanner has room for twenty term codes and that no item is indexed by more than twenty terms. The Luhn scanner matches cards in the store against the question card at a rate of 1000 per minute. Since it is a necessary characteristic of searching systems (systems in which terms are collected under items) that the total file be scanned in any search, it would require 1000 minutes or approximately 16½ hours to search a million items to answer one question. Hence a searching system even with so advanced a device as a Luhn scanner can only be used for relatively small collections or for collections which permit the division of items into mutually exclusive classes, each one of which is small enough to make searching the total class practical. The great advance of the Luhn scanner was its demonstration that free field coding could be used with punched cards and that one card could constitute the question which interrogated the store on other cards.

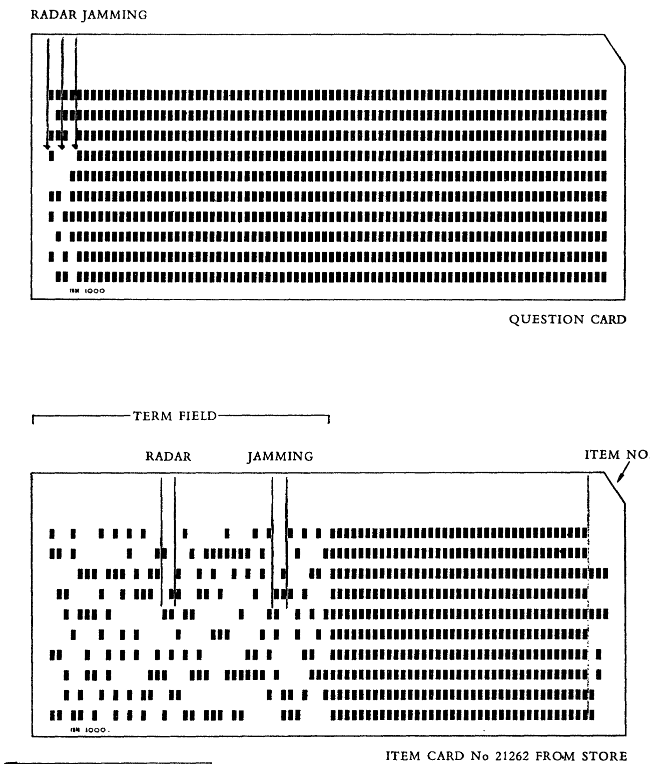


FIGURE 1

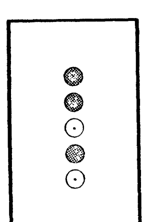


FIGURE 2. (*a*) Term code in question card.

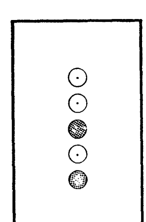


FIGURE 2 (*b*) Term code in store.

Collation

In spite of this development the inherent inefficiency of linear search has so far precluded the successful application of punched card searching to collections of any significant size which cannot be divided into mutually exclusive classes; but several relatively successful punched card installations have been organized for collating rather than searching. In setting up a system of punched cards for collating as contrasted with searching, grouping of items by terms is employed rather than grouping of terms by items. The following figure illustrates the two forms of grouping.

Searching				
1	A	M	N	O
2	B	C	D	T
3	A	B	M	R
4	L	N	O	P
5	C	G	H	K
6	F	G	M	P
7	L	P	R	T
8	H	K	L	S
9	B	C	R	S
etc.				

Collating									
A	1		3						
B		2	3						9
C		2			5				9
D		2							
F						6			
G					5	6		8	
H					5				
K					5			8	
L				4			7	8	
M	1		3			6			
N	1			4					
O	1			4					
P				4		6	7		
R			3				7		9
S								8	9
T		2					7		
etc.									

When collation is used as a matching technique, item codes collected under one term are matched against item codes collected under another term. In effect one group of item codes becomes the question which is matched against the group considered as the store. It should be apparent that collation does not require the search of the total store but only of those item

codes grouped under the terms of the question.

However, with standard collating equipment, a considerable price must be paid for this decrease in search time. A collection of 1,000,000 items indexed by an average of 20 terms would require a file of 20,000,000 cards. With 10,000 terms in the vocabulary the 20,000,000 cards would be arranged in 10,000 groups averaging 2000 cards in a group. Since a standard collator feeds 240 cards per minute from each feed, the collation of two terms (asking a two-termed question) would average between 10 and 20 minutes. This is an appreciable reduction from 16½ hours, but there are some penalties which must be faced which reduce radically the efficiency of standard collators as information searching devices.

In the first place the size of the store must be increased enormously to permit prefiling items (cards) under every term by which they are indexed, in this instance, from 1,000,000 to 20,000,000. Secondly, collators work only on arrays maintained in fixed numerical or alphabetical order. Hence, item cards must be filed (posted) to each term array and maintained in that array in a fixed order. Thirdly, cards matched by the collator and selected as answers must be refiled in proper order. If the selected cards are to be retained as an answer or are to be matched against existing groups, they may have to be duplicated so that the array from which they are selected initially can be restored to completeness for other searches.

These difficulties, which arise from the use of standard collators as information searching devices, are not attributable to the grouping of items by terms, but to the use of a collator designed primarily for interfiling of cards rather than the matching of codes. It will be seen that most of the difficulties disappear when a device like the Comac is substituted for a standard collator.

The standard collator carries out its interfiling function by noting the match, the failure to match, and the order of numbers on cards which it compares. On the basis of what the match discloses, the collator advances one deck or the other, or both (in the case in which a match occurs). The multiplication of cards from one to twenty million is not attributable to the need for additional coding space but to the fact that, since the collator reacts to a match by selecting cards or to the recognition of order by interfiling cards in proper order, it can operate on only one item code per card. But if we separate the collator's ability to match codes from the requirement for physical selection and interfiling of cards, it becomes possible to put more than one item code on a card and to signal a match by punching the matched code on another card.

The Comac

The essential function of the Comac which determines its design is simply the ability to match codes on one punched card against codes on another punched card and to punch the codes for the logical product or sum on a third card. Consider a set of item codes on card A and another set on card B. (See Fig. 3.)

The card AB can be collated with card C, etc. The final answer if it involves the product $[(A \cap B) \cap C]$ can be printed rather than punched.

It is immediately apparent that one of the features of the Comac is the fact that it does not require any refiling of selected cards into an A deck or a B deck. The degree of file reduction, however, may not be immediately apparent.

An IBM card contains 80 columns. Since, in collation, we group item codes under term codes, let us assume that 2 columns are required for the term code of each card and

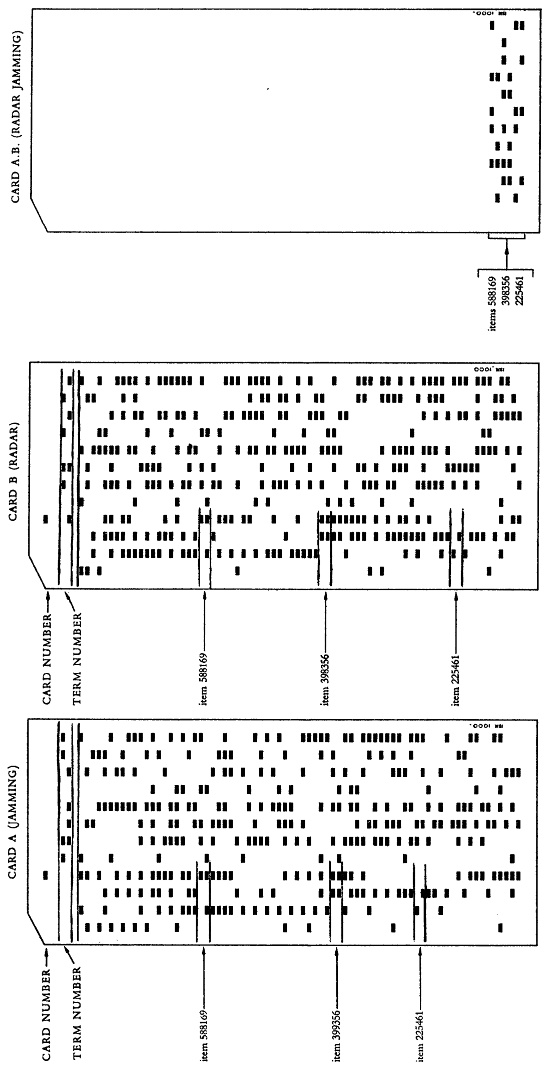

CARD A.B. (RADAR JAMMING)
items 588169
398356
225461
CARD B (RADAR)
CARD NUMBER
TERM NUMBER
item 588169
item 398356
item 225461
CARD A (JAMMING)
CARD NUMBER
TERM NUMBER
item 588169
item 399356
item 225461

FIGURE 3

3 columns for the card number. With one column blank, the remaining 74 columns can be divided into 37 two-column groups with each group containing 24 bits. The 24 bits can be divided into 6 fields of 4 giving a possibility of coding any item number from 1 to 999,999. Hence 37 item numbers ranging from 1 to 999,999 can be punched in each card.

In terms of our previous figures, namely, 1,000,000 items, 20 term codes per item, 10,000 terms in the vocabulary, the 20,000,000 item codes could be punched on 540,540 cards, providing a file reduction of 37 to 1 as compared with standard collating systems.* In more picturesque terms the card file is reduced from building size to room size. The 540,540 cards would be organized into 10,000 groups averaging 54 cards to a group. And the collating of one group against the other would involve comparing the item codes of 54 cards with the item codes on 54 cards. The groups will ordinarily not be equal and many groups will contain many more than 54 cards; but unlike the Minicard System, we can add cards to any group without dedicating space for it in the group. Hence it is appropriate to discuss the groups of a Comac system in terms of averages.

It is interesting to note that the 3,000,000 patents in the Patent Office could be handled on approximately 1,600,000 cards and given the same 10,000-word vocabulary, with an average of 162 cards in a group.

---

* This form of binary coding in columns is called "Chinese Binary" by IBM data processing people. Although it gives maximum compression of codes, it does require decimal binary converters and modification of existing punches. With regular Hollerith coding, the 37 to 1 file reduction possible with Chinese binary, drops to 14 to 1 for collections requiring 6-digit codes (up to 999,999 items) and to 18 to 1 for collections requiring 5-digit codes (up to 99,999 items).

Once the basic design requirement of the Comac is established, the Comac can operate in either of two ways depending on the amount of comparators or registers that are provided. If only one code on each card can be read at a time, both groups being collated would have to be advanced intermittently as is the case with existing collators, the difference being that the cards would be advanced two columns at a time instead of a card at a time. The Comac, which compares only one code at a time from each group, can only operate if the item codes are punched in ascending order, that is, if "code 1<code 2<code 3, etc." This is also the case with existing collators. This constraint can be avoided if enough circuitry is provided to store all the codes on a card while another card is passed over it. In such a case all codes would be matched against all codes regardless of the order of the codes on the cards. Whether or not it would be worthwhile to provide this additional circuitry would depend on the nature of the collection being indexed. If the items in the collection could not be assigned serial numbers and indexed in order, the more advanced type of Comac would be required; but if serial order could be maintained in indexing and in entering item codes on cards, the Comac which advanced and compared one code at a time from each group would be adequate.

We have not attempted in this paper to describe the Comac apparatus. However, from our studies of existing punched card equipment, binary to decimal converters, comparators, etc., it appears that once the basic concept of the Comac is accepted, the construction of a device for single code comparison represents only a very modest development effort. Actually the character of the physical equipment necessary is practically deducible from the new concept of collation as a matching and print-out process of item codes rather than a card selection and interfiling process. The development of the more advanced Comac capable of comparing multiple codes with multiple codes might require a greater

investment; but before such a development is undertaken, the value of removing the constraint of entering item codes in order should be thoroughly explored.

Advantages of the Comac System

We summarize at this point the characteristics of the Comac which make it a practical and efficient information storage and retrieval system. Some of these points have been mentioned above; others will be presented here for the first time.

1. Decrease in Time of Search

For a collection of 1,000,000 items, the grouping of item codes under term codes, as compared with term codes under item codes, reduces the time of search for a two-termed question from 16½ hours (Luhn scanner) to approximately 5 minutes. We are assuming that cards can be advanced two columns at a time and compared in the Comac at about twice the rate they can be advanced in existing card reproducers that feed cards the long way, i.e., one card every 2 to 3 seconds. Since one card in the Comac contains 37 codes, 54 cards containing 2000 codes can be read in 100 to 150 seconds. Doubling this figure to allow for the intermittent advance of two groups, we get 200 to 300 seconds or 3 to 5 minutes.

2. Multiple Access to the Store

Searching involves scanning the whole file; but collating involves comparing prefiled groups. This means that many searchers can select groups for comparing at the same time without interfering with one another. Further, it is assumed that the Comac will be so reasonable in price that any large and busy installation would have several so that searchers desiring to collate term groups would not have to queue up

at one machine. Assuming 5 minutes for the average search, five Comacs would provide an answer every one minute during a working day.

3. Elimination of Refiling

The card reproducing and print-out features of the Comac would eliminate the necessity which exists with present collators of refiling cards. That is, there would be no "return-to-normal" problem which now exists with most punched card searching and selection devices. For example, the Patent Office R & D Report No. 6 describes this problem as a major difficulty in the utilization of punched cards for searching[1].

4. Card Reproduction and Print Out

The Comac will be able to reproduce a punched card containing matched codes; but in addition it will contain a binary to decimal converter which will print out the final answer of a succession of collations, e.g., $\{[(A \cap B) \cap C] \cap D\}$.

5. Ease of Maintenance and Posting

The cumulation of item codes for punching on term cards is a simple procedure which has been worked out by Documentation Incorporated and other organizations (NSA) in connection with posting on Uniterm Cards[2].

6. Freedom from Constraints on Indexing

The reproducing features of the Comac make it fairly simple to combine terms, set up hierarchical relations between terms and the items grouped on such terms, etc. For example, a decision to establish a grouping of item codes under the general term antibiotics, that is, to collect codes previously listed under aureomycin, streptomycin, penicillin, etc., involves only changing instructions in the reproducer.

Whereas an ordinary search involves the reproduction of the logical product of the group of item codes, the collection of item codes under a general term is equivalent to reproducing the logical sum of the codes.

Hence, although it is possible to look upon the Comac as a mechanized Uniterm Index, the mechanization extends not only to the comparison of codes but to the ability to update the actual indexing and to provide any degree of order or hierarchical relationship required by the search problems confronted by the system.

References

1. Don Andrews, U.S. Patent Office, Research and Development Report No. 6, pp. 11-12.

2. Sanford and Thereault. "Problems in the Application of Uniterm Coordinate Indexing." College and Research Libraries, 17, No. 1, 19-23 (January 1956).

CHAPTER VI

THE IBM 9900 SPECIAL INDEX ANALYZER*

By R. W. Murphy**

The IBM 9900 Special Index Analyzer is IBM's version of the concept of Continuous Multiple Access Collating developed by Documentation Incorporated, Washington, D. C., under a research contract sponsored by the Air Force Office of Scientific Research, (ARDC) Directorate of Research Communications.

The IBM Special Index Analyzer is a machine designed to facilitate reference to cataloged information. It may be applied to such activities as library research, or the searching of equipment design specifications. These activities share the essential problem that, in order to make use of information which has been stored, most of it must be prevented from having to be considered by the user. If the user can specify attributes of the information in which he is interested, the Special Index Analyzer will select out of the files only those references to items of information which possess that particular association of attributes. For example, a library searching problem might be to determine all the material dealing, in the same article, with reliability, transistors, and digital computers.

---

* Reprinted with the permission of International Business Machines Corporation.

**International Business Machines Corporation, Poughkeepsie, New York.

An information retrieval system employing the Special Index Analyzer is set up on the basis that a document, or other item of information, can be categorized by a set of terms. The terms may be the names of topics, subjects, or attributes or they may be key-words actually used in documents of the collection. When items or documents are entered into the system, each is analyzed to find which terms are pertinent to it and records are made associating the item with significant items. The terms are drawn from a pre-established glossary which is used uniformly throughout the analysis of items in the collection. These records are then rearranged into a "where found" index; that is, an index consisting of subdivisions called term files, each of which includes all the references to which an individual term is pertinent. In this arrangement the term files are ready for use by the researcher, employing the IBM 9900 Special Index Analyzer to select out significant references automatically and accurately.

The operations performed by the Special Index Analyzer are of the type found in the theory of sets. These are primarily the intersection operation in which the result is the set of item references containing only those references common to the two input sets being operated upon, and the union operation which produces a resultant set containing references from either of the two operand sets. In addition, the Special Index Analyzer can perform the intersection with the various complements of the operand sets.

In its application to information retrieval the Special Index Analyzer employs, as operand sets, term files selected from the index files by the researcher. In effect, the intersection operations provide the researcher with the means of narrowing down of the scope of his search in accordance with the degree of specificity with which he can select term files. The union operation allows him to apply the narrowing down to as many terms as he feels are significant in his search.

Methods of Information Retrieval and the Application of the Special Index Analyzer

All information retrieval systems share the requirement that the content of an item or document to be included in the system must be determined by a trained analyzer and recorded. Once this has been accomplished, however, the various retrieval systems differ in the manner in which the records of information content are maintained and used. In order to characterize the methods of filing information relating terms and items, a matrix representation is most useful. In this, the set of terms constituting the glossary is arranged as one axis of the matrix, while the items of the collection are represented along the second axis. For each determination that a term relates to an item, the appropriate position of the matrix is posted with the fact of relationship. This posting of a relationship will be referred to as an "entry" and may be written either as a binary mark in the matrix or as the juxtaposition of a term code with an item code (the unit record form).

In this example, as in many practical cases, numerical codes designate both the terms of the glossary and the items of the collection. Codes which stand for terms and codes which stand for items will therefore look alike, and must be distinguished either by context or by adding a supplementary symbol.

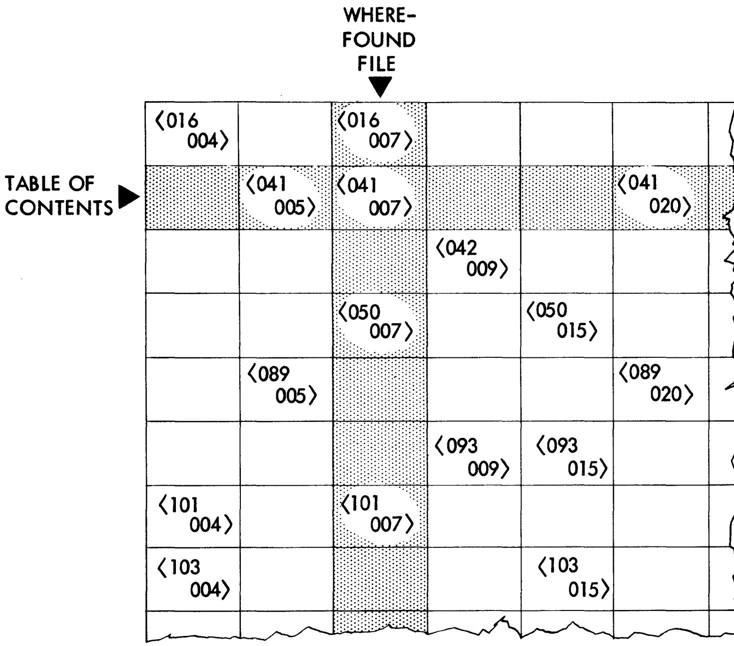
The matrix representation is intended as a conceptual device. Most practical media for the storage of data require that the information contained in the matrix be converted to a linear sequence of unit records in the process of filing. The two usual ways of linearizing the matrix are by taking successive rows, or else by taking successive columns. Each row of the matrix corresponds to a table of contents of one of the items in the library, while each column corresponds to a "where-found" listing.

— an ITEM is a physical object . . . book, document, map, record, patent . . . which is the object of the search

— a TERM is the name of a topic or attribute of the item

TERMS IN THE GLOSSARY

ITEMS IN THE LIBRARY	004	005	007	009	015	020
016	X		X			
041		X	X			X
042				X		
050			X		X	
089		X				X
093				X	X	
101	X		X			
103	X				X	



LINEARIZING BY ROW

$\{\langle 016,004\rangle \langle 016,007\rangle \ldots\}, \{\langle 041,005\rangle \langle 041,007\rangle \langle 041,020\rangle \ldots\}\ldots$

CATALOG = SET OF ALL TABLES OF CONTENTS

$= \{I_j\}$

AN ITEM FILE, $I_j = \{\langle i_j, t_k\rangle ;\ j \text{ is fixed}\}$

LINEARIZING BY COLUMN

$\{\langle 016,004\rangle, \langle 101,004\rangle, \langle 103,004\rangle \ldots\}, \{\langle 041,005\rangle, \langle 089,005\rangle \ldots\}\ldots$

INDEX = SET OF ALL WHERE-FOUND FILES

$= \{T_k\}$

A TERM FILE, $T_k = \{\langle i_j, t_k\rangle ;\ k \text{ is fixed}\}$

All the unit records made of entries to the matrix constitute a grand set representative of the entire library. The process of linearizing the matrix arrangement of the grand set results in the distinguishing of subsets which may be of either one type or another, depending on how the matrix is linearized. If the subset is taken from a row of the matrix, it contains unit records of entries all referring to the same item and designating all the terms to which the item pertains. This kind of subset is therefore a table of contents or item file, and contains term codes as the essential elements of information. The complete collection of item files is the grand set, but it is usable as the catalog of the library.

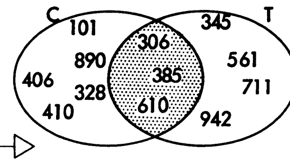
The alternate way of separating out subsets of the grand set is by taking them from columns of the matrix. Within a subset, all of the unit records will contain the same term code, but differ in the item codes. This kind of subset, or term file, tells of all the items where a particular topic, or term, is treated. The complete collection of term files constitutes an index to the library.

For information retrieval purposes, the IBM Special Index Analyzer is intended to be used with the index containing the "where-found," or term, files. Its essential function is to combine a pair of term files to produce a new term file, usually containing many fewer term codes than appeared in the original files. By means of his selection of the input term files and through his choice of the combining operations, the researcher can program the Special Index Analyzer to reduce a number of voluminous term files to just one set of item references, which meet his specifications, and at the same time, are few enough in number to allow the researcher to refer to the items directly.

The combining operations performed by the IBM Special Index Analyzer are operations on sets of item codes. There are only two basic operations which can be performed on two

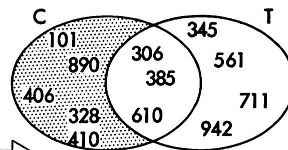
A. INTERSECTION

$T \frown C$ THE ITEM CODES COMMON TO T AND C



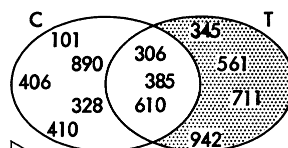
B. INTERSECTION WITH COMPLEMENT

$\bar{T} \frown C$ THE ITEM CODES IN C, BUT NOT IN T



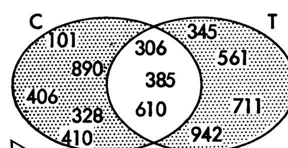
C. INTERSECTION WITH COMPLEMENT

$T \frown \bar{C}$ THE ITEM CODES IN T, BUT NOT IN C



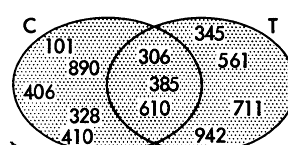
D. UNION OF COMPLEMENT INTERSECTIONS

$(\bar{T} \frown C) \smile (T \frown \bar{C})$ THE ITEM CODES IN T OR C, BUT NOT IN BOTH



E. UNION

$T \smile C$ THE ITEM CODES IN EITHER T OR C



IN THE IBM SPECIAL INDEX ANALYZER

T IS THE TERM FILE ON TAPE

C IS THE FILE BROUGHT IN FROM CARDS

operand (input) sets, the intersection ($T \cap C$) which finds the elements common to T and C, and the union ($T \cup C$) which places the elements occurring in either (or both) of T and C in a resultant set containing no duplications. In addition, set theory deals with the complement of a set ($\overline{T}$), that is, with the set containing all of the elements which are not in T. However, since any machine can only develop new sets from sets which are specifically introduced, the complement is used in conjunction with intersection in the IBM Special Index Analyzer to provide three additional operations which complete the range of operations performable on two operand sets.

Most library search operations will involve more than two term files, to be combined by means of various set-theoretic operations arranged into a program by the researcher. The program together with the term files serving as operands in the program, is equivalent to a set-theoretic expression of several variables, and may be rearranged, as the set-theoretic expression is rearranged, in order to obtain a simpler or more efficient program. Set theory provides the relations by which the expression can be reduced to the form which provides the most efficient program.

$$T_1 \cap T_1 = T_1 \qquad T_1 \cap T_2 = T_2 \cap T_1$$

$$T_1 \cup T_1 = T_1 \qquad T_1 \cup T_2 = T_2 \cup T_1$$

$$T_1 \cap (T_2 \cap T_3) = (T_1 \cap T_2) \cap T_3$$

$$T_1 \cup (T_2 \cup T_3) = (T_1 \cup T_2) \cup T_3$$

The result of taking the intersection of any number of sets depends only on what sets are involved and not on the order in which sets are combined, nor on the number of times a set is repeated. The same is true in taking the union of several sets.

$$(T_1 \cap T_2) \cup (T_1 \cap T_3) = T_1 \cap (T_2 \cup T_3)$$

$$(T_1 \cup T_2) \cap (T_1 \cup T_3) = T_1 \cup (T_2 \cap T_3)$$

$$T_1 \cap (T_1 \cup T_2) = T_1$$

$$T_1 \cup (T_1 \cap T_2) = T_1$$

$$T_1 \cap (\overline{T}_2 \cup \overline{T}_3) = T_1 \cap \overline{(T_2 \cap T_3)}$$

$$T_1 \cap (\overline{T}_2 \cap \overline{T}_3) = T_1 \cap \overline{(T_2 \cup T_3)}$$

The complement of any expression can be obtained by taking the complement of each term (letter or parenthetical term) and interchanging each cup for a cap and vice-versa.

In planning a search, the researcher will usually work out a statement in words of the course which the search is to follow. The verbal statement can then be written as a set-theoretic expression, using such symbols as T_1, T_2, to stand for the terms stated. Then, if necessary, the set-theoretic relations are used to reduce the complexity of the expression. The final step is to select the required term files out of the index and incorporate them with the program to obtain the machinable equivalent of the original statement.

STATEMENT	Retrieve the items dealing with transistors and computers but not with production, as well as the items dealing with transistors, computers, and reliability.
EXPRESSION	$(T_1 \cap T_2 \cap \overline{T}_3) \cup (T_1 \cap T_2 \cap T_4)$ where: T_1 = transistor T_2 = computer T_3 = production T_4 = reliability
SIMPLIFICATION	$(T_1 \cap T_2) \cap (\overline{T}_3 \cup T_4)$ $T_1 \cap T_2 \cap \overline{(T_3 \cap \overline{T}_4)}$
PROGRAM	1. Run in T_4 ("reliability" term file) 2. Intersection type B with T_3 ("production" term file) 3. Intersection type B with T_2 ("computer" term file) 4. Intersection type A with T_1 ("transistor" term file) 5. Print out result

IBM 9900 SPECIAL INDEX ANALYZER

Functional Characteristics of the IBM Special Index Analyzer

The IBM Special Index Analyzer is composed of three units. The first unit is a modified IBM 26 Card Punch which is used primarily for reading cards when operated with the system. It may also be employed as a standard card punch when the Special Index Analyzer is not in operation. The second unit is the logical and intermediate storage unit and contains both the control equipment and a paper tape punch and reader for retaining the intermediate results of operations. The final unit is a typewriter which is used for automatically printing the results of the search.

The Special Index Analyzer functions as a collator working with six-digit codes, rather than with complete card records. Codes within a term file are always maintained in numerical sequence allowing the Special Index Analyzer to operate upon term files containing hundreds or thousands of item codes. Item codes are read one by one from either of two inputs, one of which is the card reader and the other, the paper tape reader. After being read, a comparison is made between the two, and depending on how they compare and what operation is being performed, one or neither of the codes may be punched in paper tape and a new code brought in for the next cycle.

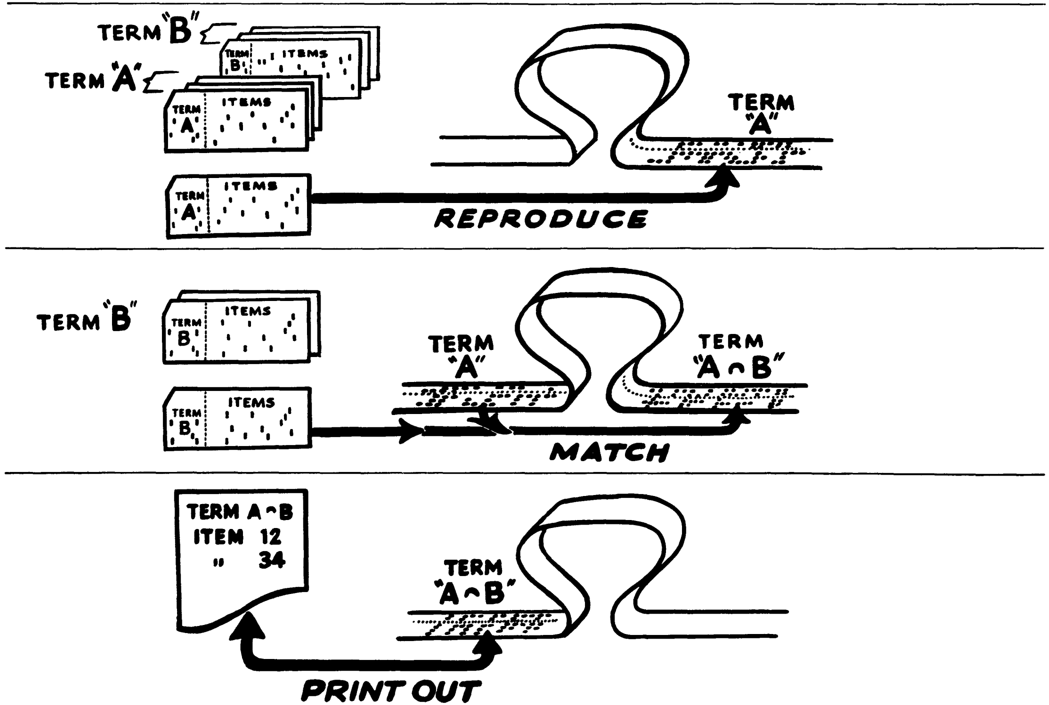
The five operations for combining sets all make use of both the card reader and the paper tape reader, and punch the resultant set into paper tape. In addition, there are certain "housekeeping" operations used for initiating a program and for reproducing the final result in a convenient form. In starting a new program, the first term file must be reproduced on paper tape before it can be combined with the second term file. After all of the term files called for by the program have been combined, the result will appear in the paper tape, from which it is usually printed out on a form.

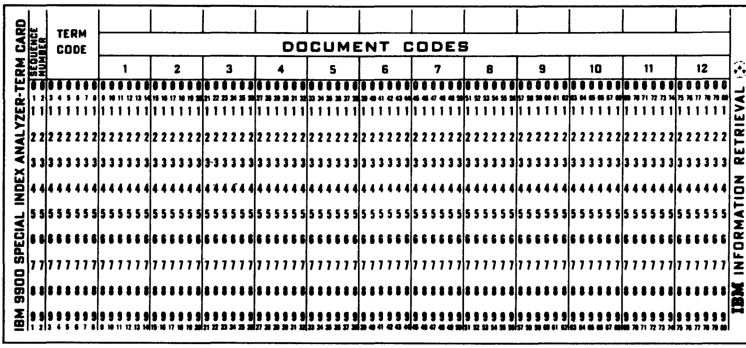
The index, or master file for information retrieval, is maintained on standard IBM punched cards. It will be composed of individual decks of cards, each deck conveying the item codes corresponding to one term code. In turn, each deck or term file will consist of one or more cards, depending on how many items are associated with that one term. Since it is necessary in setting up a program to select manually the appropriate term files from the index, punched card decks form a convenient and inexpensive storage medium. In addition, as new material is added to the library, the individual term decks can be extended to include the additional retrieval data. For this purpose, the Special Index Analyzer provides an alternative mode of operation which reproduces data from paper tape into new term cards.

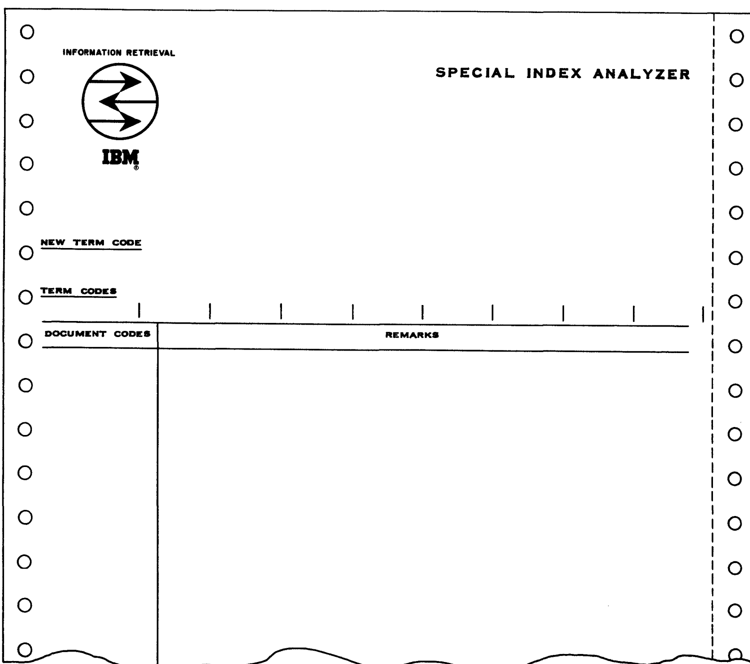
A single term card has space for thirteen six-digit codes, plus two additional digits. Of the thirteen code positions, one is reserved for the term code which identifies the term deck to which the card belongs. The two additional digit positions will customarily be used for a sequence number to locate the card within the term deck. The remaining twelve code spaces are available for item codes. These are punched in sequence from left to right across the card, with any excess positions left blank. The Special Index Analyzer recognizes the start of a new term deck by means of an X-punch in column 1 of the first card of the deck, whether the deck contains one or more cards. The X-punch does not interfere with the use of the first column for numeric data, but alphabetic punching should not be used in this column.

If it is desired to punch additional data into the card, either alphabetic or numeric, successive six-digit fields may be used. The inclusion of additional information will require that the term code and the item codes be shifted to the right, and will reduce the number of item codes that can be fitted on the card. This additional data is not processed by the Special Index Analyzer. The format employed for the term

IBM SPECIAL INDEX ANALYZER




IBM 9900 SPECIAL INDEX ANALYZER-TERM CARD
SEQUENCE NUMBER
TERM CODE
DOCUMENT CODES
1
2
3
4
5
6
7
8
9
10
11
12
IBM INFORMATION RETRIEVAL


INFORMATION RETRIEVAL
IBM
SPECIAL INDEX ANALYZER
NEW TERM CODE
TERM CODES
DOCUMENT CODES
REMARKS

cards is stored in the machine by means of a specially punched card, retained on the alternate program drum of the IBM 26 Card Punch component.

The output of the Special Index Analyzer is printed by means of a typewriter onto a form designed for convenient use by the researcher. It is important to retain a trail of the search, along with the item codes produced by the search. The Special Index Analyzer accomplishes this by first printing across the top of the page the term codes entering into the search, connected by letter symbols standing for the operations used to combine each term file with the result of previous operations. The term codes thus appear in the order in which they were used. The item codes resulting from the sequence of operations are listed in a vertical column down the left-hand edge of the page. This format provides ample room for the researcher to add further notes alongside each item code.

File Maintenance

As new items are added to the library, the index file must be extended to include retrieval data for the additions. The first step in the procedure is to assign the item code next in sequence. Then it must be determined which terms of the glossary apply to the new item. A series of cards are punched, all containing the code for the new item, and each containing the code for one of the terms determined to be applicable. The standard card form is used for this purpose, but it will normally be of a contrasting color to distinguish it from the regular cards in the file. The term code is punched in the regular field, and the item code occupies the first item field. The other positions of the card are left blank.

At this stage, each addition card must be brought together with the term deck bearing the same term code. It may be decided to bring the term file up to date immediately, in which case the term deck is removed from the file and the

updating performed. Alternatively, the addition card may be filed as such with the term deck, and the updating postponed until a suitable number of additions have been collected or until the term file is to be used in a search.

The Special Index Analyzer is used to bring the new item codes into the regular term deck. This may be accomplished with the regular union operations applied to the old term file and to each of the addition cards, which carry an X punch in Column 1 as though they were new term decks. The result is punched into blank cards using the appropriate operation. Since non-item data normally are not punched in the punch-out operation, the resulting new term deck is then gang-punched to include the term code, and punched for the sequence number.

Two variations on this procedure may be utilized if desired. To eliminate the external gang-punching of term code, the input deck may be preceded by a command card calling for the "Read in New Term Number" operation, and punched with the term code in the regular field. This term number will then be reproduced into the output deck. The second variation is intended to conserve time where very long term files are to be reproduced. By pressing a special start key, the lack of an X punch in the first term card can be ignored, so that instead of the complete old term deck, only the final card need be brought in.

New terms may also be added to the glossary and the appropriate term files built up. If a term is completely new, and appearing for the first time in new items, the previously prepared term files will be unaffected. If however the term was not previously applied to items already in the library, then all of the items where it might be significant must be reanalyzed. The most probable situation will occur where a term file has become too large and the term sub-divided into finer categories. In this case, only items in the term

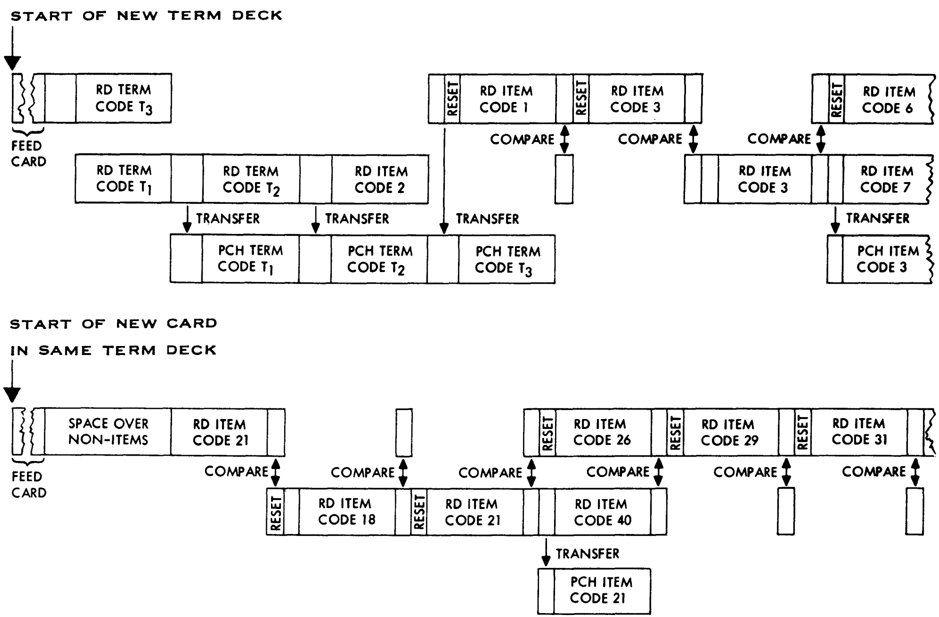
file in question must be reinvestigated. It may also be desirable, in the event that certain combinations of term files are repeatedly employed, to add the combination to the index under a new term code standing for the combination. This action adds no new information to the retrieval index, but may add to the convenience or speed of searching.

Speed of Operation

In its usual operation, the Special Index Analyzer will be working with term files of various lengths, ranging from tens of items to thousands. The intersection of two term files may also contain a wide range of items, from a small fraction of one of the original files to the entirety of the smaller original file. Thus the operating times experienced in practice will follow a statistical distribution around average values which are typical for the installation.

The basic speed of the Special Index Analyzer is eighteen cycles per second. A cycle may consist of reading or punching a character in paper tape or card, typing a character, or performing certain internal operations. In general, each item code of six characters requires eight cycles (0.444 sec.), including six cycles for reading into a register, one cycle for comparison and one for resetting or for transferring the contents of the register. Each of the two inputs and the output channel makes use of a separate register for storing the term or item code and all three can operate concurrently. However, unless the two input registers contain equal codes, only one of the inputs will be actuated prior to processing the next code. The major portion of the time required to perform one pass will be due to the processing of item codes. However, there will also be contributions due to the reproduction of term codes at the start of the pass, and due to the feeding up to the item code portion of each succeeding card in the term deck.

TIMING OF AN INTERSECTION OPERATION



TIME FOR ONE INTERSECTION OPERATION

$$\cong 0.25 + \frac{8}{18}i + \frac{l_i}{12} + \left(l_i + l_{ni-1} - l_{ni}\right)\frac{8}{18} \text{ seconds}$$

FEED 1ST CARD

REPRODUCE TERM CODES

FEED + SPACE OVER NON-ITEM DATA IN SUCCEEDING CARDS

"HOUSEKEEPING" STEPS COMMON TO ALL COMBINING OPERATIONS

WHERE:

i = NUMBER OF THE COMBINED TERMS

l_i = NO. OF ITEM CODES IN THE i^{th} TERM DECK

l_{ni-1} = NO. OF ITEM CODES ALREADY ON TAPE

l_{ni} = NO. OF ITEM CODES IN RESULT TAPE

For determining the time required to perform an entire program of several combining operations, it is easiest to determine the total number of distinct (non-overlapping) readings or punchings of item codes. Once this total has been obtained it is multiplied by 0.444 seconds and added to the time required by the housekeeping steps. In all complete programs, there will be in addition to the combining of term files, an initial stage of reproducing the first term deck on paper tape, as well as a final stage of printing out the results.

A program of intersections only is probably the most common type and is therefore most indicative of machine performance. For this type of program, the number of distinct input-output steps is obtained by writing the number involved in each stage:

1st. stage - reproduce T_1 on tape	l_1
2nd. stage - form $T_1 \cap T_2$	$l_1 + l_2 - l_{\cap 2}$
3rd. stage - form $T_1 \cap T_2 \cap T_3$	$l_3 + l_{\cap 2} - l_{\cap 3}$
rth. stage - form $T_1 \cap \cdots \cap T_r$	$l_r + l_{\cap r-1} - l_{\cap r}$
$(r+1)$st. stage - print $T_1 \cap \cdots \cap T_r$	$l_{\cap r}$
	$l_1 + \sum_{i=1}^{i=r} l_i$

Thus, for a program of intersections only, the time required depends upon the total number of item codes involved and upon which term file is chosen to be first, and is independent of the numbers of item codes in the intermediate and final intersections.

This expression presupposes that the reading of paper tape is being performed sufficiently in advance of punching to allow the minimum length of four inches of tape between the stations. If the data to be punched in tape, including both terms and items should be less than six codes, then blanks will be inserted to keep the length of tape at the minimum.

These blanks will require additional time to be read on the next pass; however the total time added will, in general, be very small.

The total time required to perform a program of intersections of r term files thus becomes:

$$0.22\,r^2 + 0.47r + 0.44\,\ell_1 + 0.52\sum_{i=1}^{i=r} \ell_i$$

The time required to perform any one of the other set-theoretic operations is exactly the same as for performing the intersection operation. This is because the two input term files are read separately, except when item codes are equal, or in other words, except for the item codes belonging to the intersection. Punching of the result into paper tape always overlaps a reading of one or both inputs, and does not contribute to the time. However, the number of item codes in the intersection of the two operand sets is not generally known beforehand, and therefore the size of the result set from these other operations cannot be predetermined. Unlike the case of the intersection operation performed successively in a program, the size factor does not cancel, but remains significant in programs utilizing these other operations.

An upper bound can be determined for these more complex programs, if it is assumed that there are no item codes common to the intermediate intersections of term files. This assumption is equivalent to assuming that the size of the intermediate term file is the sum of the sizes of the two component term files for the two operations, "union" and "union of complement intersections", or that it is the same size as the non-complemented term file for the two "intersection with complement" operations.

In the case of a program consisting of all union operations for combining term files, the upper bound of the number of distinct input-output operations is:

1st. stage - reproduce T_1	ℓ_1
2nd. stage - form $T_1 \cup T_2$	$\ell_1 + \ell_2$
3rd. stage - form $T_1 \cup T_2 \cup T_3$	$\ell_1 + \ell_2 + \ell_3$
rth. stage - form $T_1 \cup \cdots \cup T_r$	$\ell_1 + \cdots\cdots + \ell_r$
$(r+1)$ st. stage - print out $T_1 \cup \cdots \cup T_r$	$\ell_1 + \cdots\cdots + \ell_r$
	$\sum_{i=1}^{i=r} (r+2-i)\ell_i$

Programming the IBM Special Index Analyzer

Programming the Special Index Analyzer consists of arranging a sequence of term decks to serve as the input data, and then, at the start of reading each term deck, calling for a particular operation. Two modes of programming the machine are provided, the automatic mode and the manual mode. The automatic mode makes use of command cards in addition to the term cards, and allows the entire search procedure to continue without operator attention, except possibly for reloading the feed hopper and emptying the stacker.

In general, the manual mode will cause the Special Index Analyzer to stop after reading each term deck. The operator causes the machine to continue with the next operation by pressing the appropriate operation button. If the operation is one of the combining types, the Special Index Analyzer will only continue if there is a term deck in the field hopper. The type-out can be called for whether or not there are cards in the hopper. However, if it should be desired to type out an intermediate result and then continue with a combining operation, the paper tape must be repositioned by the operator, so that the last punched term file can be read again in the continuation of the program.

The automatic mode of operation permits as extensive a search as desired to be performed entirely automatically after the term decks have been assembled, placed in the hopper, and the "Start" button pushed. The automatic mode may be used either with or without command cards. A command card is used preceding each term deck for which a combining operation different from the previous is wanted. If no command card is used, the next term deck is combined with no change of operation type. Type-out, however, is not done automatically unless there is a type-out card in the hopper. Thus, if hopper capacity is insufficient for a set of term decks, the machine will stop and await further action by the operator.

Rule for Timing a Typical Search

The total time T required to perform a program of intersections of r term decks is:

Total time, $T = 0.22 r^2 + 0.47 r + 0.44 \ell_1 + 0.53 \sum_{i=1}^{i=r} \ell_i$

Where:

r = number of term decks

ℓ_1 = number of items in the i^{th} term deck

i = 1, 2, 3, etc. depending upon whether it is the 1st, 2nd, or 3rd, etc. term decks

ℓ_i = number of items in the i^{th} term deck

Following is an example of the calculation of time required for the intersection of three term decks.

Terms	Deck	Cards per Deck		Items per Deck
Reliability	second	25	(25·12)=	300
Transistors	first	20	(20·12)=	240
Digital Computers	third	30	(30·12)=	360
		75 Cards		900 Items

Result of typical intersection:

$$T = (0.22 \cdot 9) + (0.47 \cdot 3) + (0.44 \cdot 240) + (0.53 \cdot 900)$$
$$1.98 + 1.41 + 105.60 + 477.00 = 585.99 \text{ sec.}$$

$$\frac{585.99 \text{ seconds}}{60 \text{ seconds}} = 9.77,$$ or 10 minutes required machine time for the intersection of three term decks.

CODE	OPERATION	RESULT SET	TYPE-OUT SYMBOL	DESCRIPTION
0	TYPE AND PUNCH			Type out term and item codes from tape, and punch item codes into term codes. Punch new term code if previously programmed.
1	READ IN NEW TERM CODE			Read in new term code from command card in preparation for final punch-out operation (codes 0 or 7)
2	INTERSECTION	$T \cap C$	A	Put item codes common to term file from tape and to term file from cards into tape.
3	INTERSECTION WITH COMPLEMENT FROM TAPE	$\bar{T} \cap C$	B	Put into tape item codes from cards, provided they are not read from tape previously prepared.
4	INTERSECTION WITH COMPLEMENT FROM CARDS	$T \cap \bar{C}$	C	Put into tape item codes not on cards but in tape.
5	UNION OF COMPLEMENT INTERSECTIONS	$(\bar{T} \cap C) \cup (T \cap \bar{C})$	D	Put into tape item codes not common to tape and cards.
6	TYPE OUT			Type out last-punched paper tape in standard format.
7	PUNCH OUT			Punch out last-punched item codes from tape in term card format. Punch new term code if previously programmed.
8	UNION	$T \cup C$	E	Put into tape item codes appearing in either or both tape and cards.

CHAPTER VII

THE IBM UNIVERSAL CARD SCANNER FOR PUNCHED CARD INFORMATION SEARCHING SYSTEMS*

By H. P. Luhn**

Introduction

It is typical of mechanical information scanning systems that the discovery of wanted information is substantially a problem of comparing a given code pattern with the various code patterns contained in a collection of stored records. When comparing or matching, there is no need for an information processing machine to interpret the significance of the code patterns as such, as is essentially the case in computing machines and associated devices. Instead, it is only necessary to establish the coincidence of a set of code elements or marks as contained in a given pattern on the one hand and any of the stored patterns on the other. Whenever such coincidence occurs, the only basic requirement is that the record which caused the match be appropriately identified.

Since punched cards furnish a convenient record storage medium, their use in information searching systems has

---

*Reprinted with the permission of International Business Machines Corporation.

**International Business Machines Research Center, Yorktown Heights, New York.

many attractive features. Information patterns of a great variety of coding schemes may be recorded by punched holes; also, punched card scanning devices can perform matching operations by comparatively simple methods.

An electronic machine that answers the requirements of information retrieval by scanning of punched cards has been developed by IBM. This machine, called the "Universal Card Scanner" (UCS), scans cards fed through it in a manner similar to that employed on conventional punched card sorters. It is capable of discovering whether any one or several of a given set of patterns are wholly or partly contained in any of the record cards scanned. This function is performed by a "no-pulse matching" process under the control of a "question card" which contains prototypes of the patterns sought, likewise represented by punched holes. This is the adaptation of an electronic method to the optical principle of "matching by black-out," employed in an earlier experimental IBM card scanning machine, frequently referred to as the "Luhn Scanner." As was the case in the earlier model, the present machine features the use of a punched IBM card (Question Card) for furnishing the patterns to be searched for in a record file.

The particular matching process employed in the UCS requires that the pattern on the record cards be given in complementary form, i. e., the various marks or elements of the pattern need to be represented by the absence of holes and all else by the presence of holes. Wherever this method of recording is impractical, the machine may be conditioned to obtain the effect of such inverted patterns, by electrical means, from normal recordings.

The identification of records which answer given patterns is brought about by physical segregation of the affected record cards from the rest of the file. Such separation is performed by diverting the affected cards to a separate pocket

in the machine. However, additional pockets are provided to permit the grading of the responding cards in accordance with certain conditions that may be set up on the control panel of the machine. If, for instance, the question is made up of several individual patterns, the responding cards may be distributed into several pockets in accordance with the number of individual patterns that were matched in each case or in accordance with some other criteria of classification.

A more detailed description of the features of the machine and of its operation follows.

The Record Card

The record cards for searching by the UCS have the form of standard IBM punched cards. Because of the particular manner in which these cards are processed by the machine, the patterns to be scanned have to conform with certain requirements as to location and arrangement on the card. Basically, the machine scans a card as a unit, i. e., whatever is contained within the twelve positions of the card columns is treated as one continuous pattern and a match or lack of match is determined once per card on the basis of such twelve position patterns. Patterns may be of any width desired and a plurality of them may be recorded across the card at predetermined locations, either adjoining or overlapping each other.

There are many types of coding that may be used to create searchable patterns for scanning by the UCS. It will depend on the individual requirements of a system as to which of the many coding schemes is most effective and no attempt is made here to evaluate their respective merits.

The simplest pattern is that produced by alpha-numeric Hollerith codes in conventional fields across IBM punched

cards, where each field stands for a predetermined class of information or data. The UCS is capable of detecting the presence in such cards of given data in given fields, not just singly, but in a plurality of such fields concurrently and in any of the many combinations that an inquiry may demand. The same is true for information recorded discretely in fields by means of codes other than Hollerith. In all of these cases the condition of a match is the one-to-one agreement between the given patterns and the patterns that may be contained in the record cards.

The usefulness of searching devices has been appreciably extended by the realization that more than one item of information may be entered in the same field if the ambiguity created by the resulting patterns can be tolerated and be held within practical limits. In patterns created in this fashion it becomes the function of a searching device to discover the inclusion of a given pattern within the patterns that may be contained in the record cards. The UCS is capable in this case of detecting the presence of any one or several given patterns within a common field and, again, in any combination desired.

The punching by Hollerith code of several entries in the same field is the simplest form of such superimposed recording although its practicability is limited in the case of numeric data. The use of alphabetic information, such as open-language words, on the other hand, is more practical because of the redundancy inherent in word spelling. The great advantage here is the fact that such words need not be translated into code words but may be used in their original form. If many entries are required, the crowding of a common field may be avoided by spreading the words over a larger portion of the card. A very simple method of accomplishing this is to assign a certain column as starting point of a word in accordance with the starting letter or letters of that word. Randomness of distribution of recorded

words over the field may be improved by making allowance for the frequency of occurrence of starting letters. (See table, Figure 1.) Also, since the starting letter or letters are identified by the starting column, they may be omitted from the spelling in the field.

This principle of staggered recording makes it possible to express specific relationships of words in terms of word pairs, triplets, etc. This is a very desirable and useful property of coding systems. By having an associated word adjoin the principal word, it assumes a location different from the one it might have had if entered by itself. Since this abnormal location is a function of the location of the principal word the probability of improper association of word groups is substantially eliminated.

In many information searching applications it is necessary to devote a certain portion of the record card to the identification, in machine-readable form, of the record itself, thereby reducing the space that can be given over to the representation of search patterns. More compact methods of coding must therefore be employed to obtain comparable resolution. The use of fixed fields of appropriate size for entering dispersed code marks is an established method, commonly known as "superimposed coding". There are several ways of accomplishing this kind of coding, involving dictionary look-up of pre-assigned codes or the systematic derivation of codes from the original words. Since the effectiveness of this coding method may be improved by insuring randomness of distribution of code marks in a field, the codes are often randomly distributed by means of random numbers or by the use of randomizing squares*. The

* H. P. Luhn, "Superimposed Coding with the Aid of Randomizing Squares for Use in Mechanical Information Searching Systems," 1956, IBM Product Development Laboratory, Poughkeepsie, N. Y.

IBM UNIVERSAL CARD SCANNER

Column assignment of starting letters and numbers for scattered entry of a plurality of open language words in a 72 column card field by means of Hollerith Code. (This distribution of starting letters is based on Webster's New Collegiate Dictionary.)

Col.	Col.	Col.	Col.
1 AA-AK	19 GA-GN	37 PO-PRE	55 WI-WY
2 AL-AN	20 GO-GY	38 PRI-PY	56 X
3 AO-AZ	21 HA-HE	39 Q	57 Y
4 BA-BE	22 HI-HY	40 RA-REF	58 Z
5 BH-BO	23 IA-INF	41 REG-REY	59 1
6 BR-BY	24 ING-IZ	42 RH-RY	60 2
7 CA	25 J	43 SA-SC	61 3
8 CE-CI	26 K	44 SE-SH	62 4
9 CL-COM	27 LA-LE	45 SI-SN	63 5
10 CON	28 LI-LY	46 SO-SQ	64 6
11 CR-CZ	29 MA	47 ST	65 7
12 DA-DE	30 ME-MN	48 SU-SY	66 8
13 DH-DI	31 MO-MY	49 TA-TE	67 9
14 DO-DY	32 N	50 TH-TO	68 0
15 EA-EN	33 O	51 TR-TZ	69
16 EO-EZ	34 PA	52 U	70
17 FA-FJ	35 PE-PH	53 V	71
18 FL-FY	36 PI-PN	54 WA-WH	72

Note: Since the first letter of a word or word combination is identified by the starting column, it need not be punched. Therefore, the recording of a word may begin by punching its second letter in the column designated by the initial letter or letters. Example: ARITHMETIC - start by punching 'R' in column 3, CONSTANT - start by punching 'O' in column 10.

FIG. 1

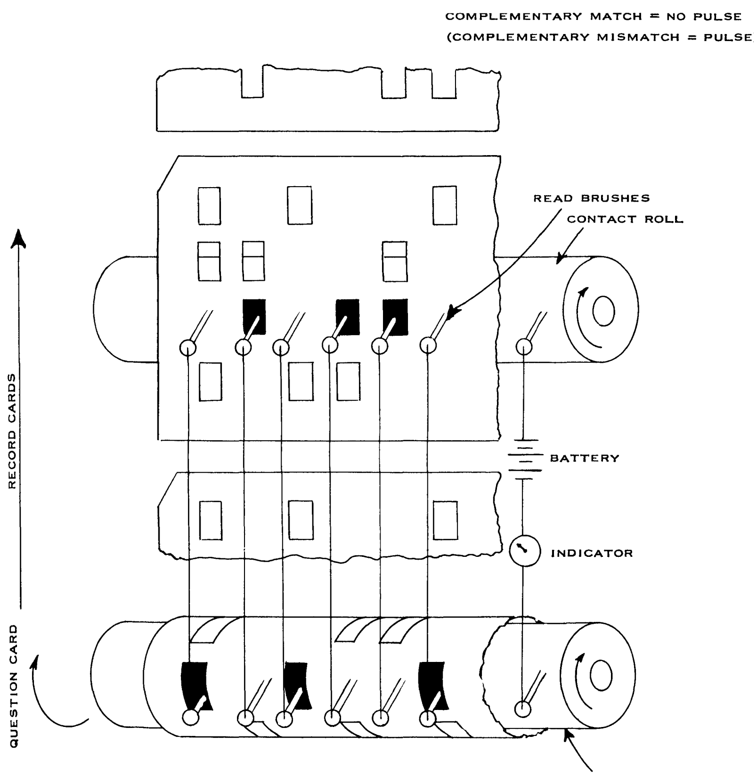
UCS is capable of discovering the inclusion of a plurality of given codes of this kind in the patterns of record cards and under a variety of conditions that may be desired.

Depending on the objectives of searches, several such coding fields may be employed in order to break down the coded information by classes. In whichever way the information is represented on the record card by means of patterns of the types just reviewed, it is necessary to know the location in which a given word or its code may be found when scanning the record cards.

The Coding of Questions by Means of the Question Card

The feeding device of the UCS has access to all 80 columns of a record card by way of 80 read-brushes. The advance of record cards through this feeding device is paralleled by the advance of the question card, wrapped around a rotating cylinder, and passing by a set of read-brushes in synchronism with the reading of a record card. The differential time signals picked up by any of the record card brushes may therefore be compared with the differential time signals picked up by any of the question card brushes in such a manner that if a columnar pattern represented by punched holes in the question card is contained in the columnar pattern represented in the record card, no output signal is emitted by the associated matching device. In all other cases a pulse is emitted. It is therefore possible to couple any number of such individual matching devices in parallel to analyze a corresponding number of columns of the record card and to discover agreement between a pattern on the question card with a pattern on the record card through the absence of pulses during the full rotation of the question card. Disagreement in any of the columns of any of the differential time pulses will cause a pulse to be emitted, signaling a mismatch. The principle of no-pulse matching is illustrated by diagram, Figure 2.

PRINCIPLE OF NO-PULSE MATCHING OF PATTERNS ON QUESTION CARD WITH PATTERNS ON RECORD CARDS



NOTE: The above scheme demands that record cards be punched in complementary fashion. When using superimposed coding methods this requirement complicates the creation of records and provisions have therefore been made in the UCS for reading record cards in complementary fashion by electrical inversion, if so desired.

Figure 2

The matching devices may be coupled to form several groups so that several patterns may be analyzed independently. The results of either a match or mismatch for each of the patterns are momentarily stored so that they may be tested for the fulfillment of the conditions stipulated for the particular search, as will be described in detail in a subsequent chapter.

Because of the above matching process, the preparation of a question card is quite simple. One type of question card cylinder provides for the coding of six question terms of up to twelve columnar patterns each. Consequently the question card is divided into six fields of twelve columns each. Each question term may be punched into one of the fields for individual matching. If it is desired to search for certain terms under the condition that all must be present to fulfill the question, then all of them may be punched into the same field, thereby gaining space for a corresponding number of additional question terms.

Alignment of Record Card Patterns with Question Card Patterns

By means of pluggable connections on the control panel of the Scanner any column of the record card patterns may be associated with any one or several columns of the question card patterns. This facility permits the simultaneous comparison of a number of individual patterns, recorded in individual fields on the question card, with a single pattern on the record card. It furthermore makes it possible to analyze any size pattern on the record card by way of a single field of the question card as long as the elements of the code combination searched for are located in not more than twelve of the several columns comprising the pattern on the record cards. Diagram Figure 3 shows some typical alignments of record card columns with columns of the several fields of the question card.

ALIGNMENT OF RECORD CARD PATTERNS WITH QUESTION CARD PATTERNS

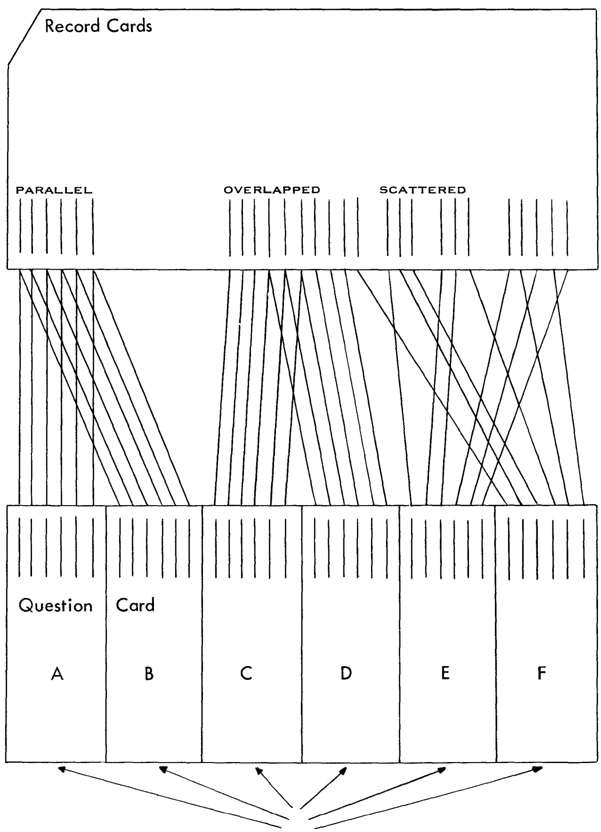


Figure 3

This versatility of association between columns of the question card and columns of the record card gives the system designer a great deal of freedom in the manner in which coded information may be organized and arranged for greatest efficiency of searching.

Conditioning of Questions

As was indicated before, the incident of a match or of a mismatch of each question term is momentarily stored in order to establish whether the combined results of a comparison satisfy the conditions stipulated for the search. Whatever their form may be, they are expressible by means of some algebra of logic, such as Boolean Algebra, which operates in terms of "and," "or," and "not."

In order to facilitate the conditioning of questions with the aid of Boolean Algebra, all permutations of six question terms are available through outputs on the control panel. With the aid of this arrangement a wanted condition may be programmed by simply plugging into the appropriate one or several of 64 plug hubs of the control panel and of using these outputs to control the distribution of the record cards into certain pockets of the machine.

Amongst the many conditions of search that may be programmed, the UCS features a searching procedure based on a statistical count of the number of question terms matched by each record card. By this method, cards are "spread" over a number of pockets in accordance with the number of matches discovered in each. Cards which did not match are deposited in pocket 0, cards which scored 1 match are deposited in pocket 1, cards which scored 2 matches are deposited in pocket 2, and so forth. This method of graduating the responses becomes particularly effective when the number of question terms is large or where the formulation of logical statements becomes too complex.

In those cases where the number of question terms exceeds the capacity of the question card and of the comparison devices, additional scanning runs must be performed. In order to extend the statistical count to become cumulative over the various runs, the attendant complexity of card manipulation is resolved by an automatic pocket shifting device. This device is controlled by special cards placed into the various pockets after each run.

Inversion of Reading Signals

A further condition which is subject to programming concerns the signals which are picked up by the read-brushes of the record card feed. As was pointed out earlier, the method of detecting coincidence of pattern elements is based on no-pulse matching. Ideally this method requires that the patterns carried by the record cards be represented by the absence of holes. This form of representation is impractical in certain cases and means have therefore been incorporated in the UCS to permit the scanning of patterns represented by holes at the option of the user through appropriate plugging on the control panel.

Because of the facility of the UCS of inverting or not inverting the signals from any of the read-brushes, its versatility has been significantly expanded. This means that the designer of searching systems has an additional degree of freedom and may avail himself of those of a variety of coding schemes which are best suited for representing and searching a particular class of information.

The ability of testing for the presence or absence of holes may be utilized to ascertain whether in a given column of a record card certain holes are present while certain others are absent. In this case the signals from the affected record card read brush are directed to two columns of a question card section, one column containing a hole pattern for the

holes to be present and the other containing a hole pattern for the holes to be absent in the record card column. Through appropriate plug connections the read brush signals directed to the first column are inverted while those directed to the second column are left unchanged. If the combined output results in the absence of pulses, a match is indicated.

APPENDIX

EXAMPLES OF INFORMATION RETRIEVAL SYSTEMS USING THE UNIVERSAL CARD SCANNER

By way of example, two experimental Information Retrieval Systems, implemented by the Universal Card Scanner, are described on the following pages:

Example 1: IBM Oswego Library Project
Example 2: Bibliography Project on Information Retrieval and Machine Translation Literature

Example 1: IBM Oswego Library Project

This experimental system was recently developed and put in operation at the technical library of the IBM Military Products Division at Oswego, New York, through the efforts of Messrs. C. Kuljian and D. Marr. The system uses a standard card catalog method of filing technical reports by title and author in a conventional manner. Mechanization is applied to the subject classification phase of the system and the retrieval of information by means of manually assigned key terms.

Encoding of Documents by means of a special thesaurus

The assignment of key terms is carried out with the aid of a special thesaurus. The application of a Roget type thesaurus to the problem of encoding was first introduced by H. P. Luhn in 1952* and in 1954 the encoding of 1200 technical reports by means of a thesaurus was carried out at the IBM

* H. P. Luhn, "A New Method of Recording and Searching Information," American Documentation, January, 1953.

Technical Library at Endicott, New York, by Mr. D. S. Tompkins and his staff. Subsequently this special thesaurus has been adopted for the Oswego system and has been simplified to reflect particular local interests.

The thesaurus consists of a set of categories and an index or dictionary. The categories have been established in accordance with the notions that are typical of the technological field of interest which the library is to serve and support. Each notional category is characterized by a definition and is identified by a three-letter code. The various words or technical terms which have been assigned to a particular category are recorded under the heading of the code word by way of a punched card dictionary file, maintained in alphabetical order. Certain technical terms are defined by a combination of several notions and therefore identified by a chain of the applicable code words. In this case a technical term is recorded under each of the several code words constituting the composite code. A print-out of this thesaurus file serves as reference when indexing a new document. A sample page of this type of thesaurus is shown in Figure 1.

In order to facilitate reference to the thesaurus for the assignment of appropriate notional codes when encoding a new document, an alphabetic index or dictionary is also compiled, listing all the words, together with a definition and the code of the one or several notional categories they have been assigned to. A sample of this dictionary is given in Figure 2.

The three-letter code words used in connection with the thesaurus have been selected from the list of self-demarcating code words compiled and published by IBM in 1953. * In choosing particular codes for given notions a reasonable effort was made to apply mnemonic principles to facilitate

---

* H. P. Luhn, "Self-Demarcating Code Words," IBM Engineering Laboratory, Poughkeepsie, New York, 1953.

memorizing of the code word assignments.

Preparation of Record Cards for Mechanical Searching. (See diagram, Figure 3)

A record card file has been assembled which is to permit the retrieval of information through characterization of an inquiry by means of one or several of the thesaurus categories. In order to accomplish this, the code words assigned to a given document were entered by super-imposition into a 12 x 12 position common field of a punched card. The remainder of the card was given over to the recording of the title of the document and of its reference number. Where necessary, abbreviations were used to fit titles into the available space, in accordance with the requirement of the system that each document be represented by a single record card.

The entry of code words into the 12 x 12 field was carried out with the use of "randomizing squares," based on methods more fully described in a separate paper. * The particular method employed here consists of spelling code words in terms of successive letter pairs. These letter pairs are marked as the intersections of rows and columns, where a particular row stands for the first letter of a pair and a particular column for the second letter. The assignment of letters to the 12 rows and columns is shown in Figure 4. The specific method of spelling used here is referred to as "chain spelling" and consists of linking the pairs by repeating the second letter of a pair as the first letter of the succeeding pair. This chain is closed on itself by forming an additional pair by "end-around spelling" of the last letter and the first

* H. P. Luhn, "Superimposed Coding with the Aid of Randomizing Squares for Use in Mechanical Information Searching Systems," IBM Product Development Laboratory, Poughkeepsie, New York, 1956.

letter of the affected code word. For instance, the code word TUG is spelled TU, UG, GT. The spelling of this word and of the word DEV has been indicated in the randomizing square of Figure 4 by marks x and o respectively. If it is desired to indicate that TUG and DEV form a composite code word, this could have been accomplished by the spelling TU, UG, GD, DE, EV, VT. This modification permits the recording of explicit relationships, a most useful function in retrieval schemes, while at the same time permitting identification of the individual words by disregarding end-around letter pairs.

The patterns representing the various code words have been derived manually and punched into an appropriate field of the dictionary cards at the time of preparation of the dictionary. The use of IBM Port-a-Punch prescored cards greatly facilitates the manual creation of these dictionary cards. The preparation of the record card patterns involved the reproduction and superimposition of the affected individual dictionary patterns into the common field of the record card.

In order to facilitate the preparation of record cards as well as of question cards for searching by means of the Universal Card Scanner, the location of the randomized pattern fields, in the dictionary and record cards, conforms with the field locations on the question card. This permits the use of standard IBM card punch (such as 24, 26) for the preparation of record and question cards by simple duplication from the applicable dictionary cards.

The record card is arranged to carry three 12 x 12 fields, identified by 1, 2, 3, and located in the left half of the card. The right half of the card is used for identification of the document recorded on the card. This identification may consist of an abbreviated title and of the document serial number. This layout of the record card is shown in Figure 6. In the

present system only field 1 is used and the area occupied by fields 2 and 3 has been utilized to extend the recording area devoted to identification of the document.

The dictionary card resembles the record card in that it too has three 12 x 12 fields in its left half, designated by A, B, C. The right half serves the recording of the dictionary term, its code and a card serial number. The dictionary term pattern is entered on this card in triplicate, i.e., the identical pattern is recorded in fields A, B, C. The layout of the front of the dictionary card is shown in Figure 8. When it is desired to enter the pattern of a given dictionary card into the record card by duplication on a card punch, such duplication may then be made selectively in any of the three locations on the record card.

<u>Preparation of a Question Card for Search on the Universal Card Scanner</u> (See diagram, Figure 5)

The Universal Card Scanner used for the present system is arranged for setting up six individual query patterns of 12 columns each. The question card is therefore divided into six sections of 12 columns each, as illustrated by Figure 7. When a particular question has been formulated, the problem arises of recording the patterns of the selected dictionary terms into the proper ones of the six fields of the question card. This is readily accomplished in a manner similar to that employed for preparing the record cards. However, since the question card requires the recording of patterns in the three additional fields D, E, and F, the duplication into these areas is accomplished by turning the dictionary card around, with its former right edge pointing to the left, (see Figure 9), and by reproducing the mirror images into any of the fields D, E, or F of the question card. The mirror effect is compensated for by appropriate wiring on the control panel of the scanner.

In those cases where composite code words are recorded by chain-spelling, the un-coupling of such words calls for the manual creation of the desired individual patterns on the question card.

Scanning Operation and Selection

The formulation of the question may include certain specific conditions under which the selection of documents shall be carried out. In the present system preference is given to the initial segregation of cards into separate pockets in accordance with the total number of question terms matched. If the resulting selection produces an unusual number of cards, then a more specific set of conditions may be programmed on the control panel and the cards just selected would be submitted to another scanning.

The cards finally selected are printed out on a tabulator (such as the IBM 407) in the descending order of matches scored or some other appropriate sequence. The resulting printed bibliography, containing title and document number, is then delivered to the inquirer.

THESAURUS

Word	Definition	Code
Abrasion	To wear away	BAB
Brush	To wear away	BAB
Chafe	To wear away	BAB
Erode	A wearing away	BAB
File	To wear away	BAB
Fray	To wear away	BAB
Fret	To wear away	BAB
Galling	Wearing away	BAB
Grate	To wear away	BAB
Mill	To wear away	BAB
Rasp	To wear away	BAB
Rub	To wear away	BAB
Score	To wear away	BAB
Scrape	To wear away	BAB
Scratch	To wear away	BAB
Scrub	To wear away	BAB
Smear	To wear away	BAB
Tribo	Rubbing Friction	BABGOP
Triboelectricity		BABGOPXED
Approach	Way or means of approach	BAC

FIG. 1

DICTIONARY

Word	Definition	Code
Acquisition	Act of acquiring	QUX
Acrylic	Thermoplastic resin	KIC
Action	The doing of something	DUZ
Activate	To produce an effect or result	MEN
Active	Producing an effect or result	MEN
Actuate	Cause to act	TUG
Actuator	Operating device	TUGDEV
Acute	Of delicate composition	NIC
Adapt	Adjusting to conditions	FIX
Adaptor	Joining device	YOKDEV
Add	To increase	BAF
Adder	A means of collecting or gathering	BAFDEV
Addition	Increase or addition	GAN
Address	Destination, identification	DES
Adhesion	A sticking to	GOP
Adhesive	Substance for sticking together	HES
Adiabatic	No gain or loss in heat	NOG
Adjoint	An addition	JUN
Adjunct	An addition	JUN
Adjust	To set right	FIX

FIG. 2

PREPARATION OF RECORD CARD

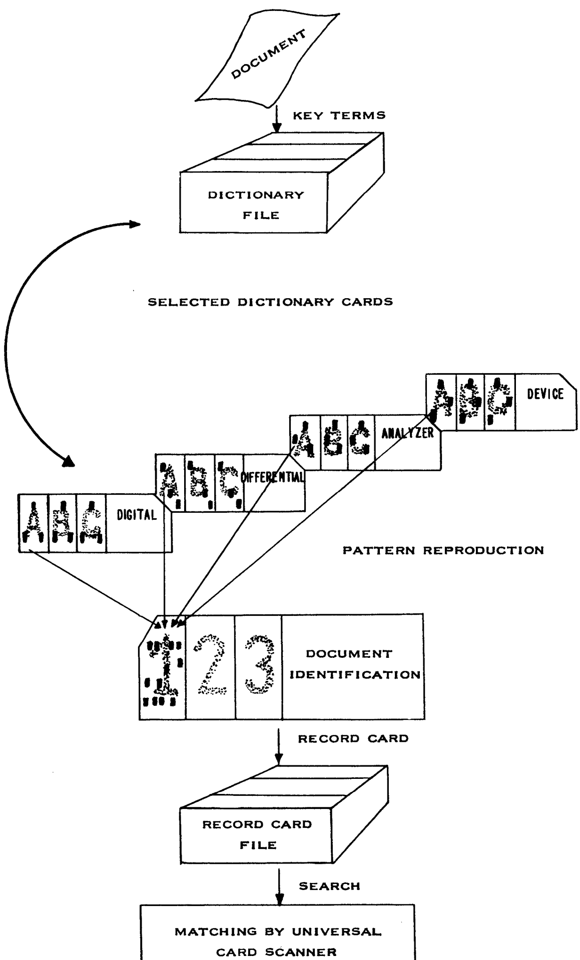


Figure 3

RANDOMIZING SQUARE

	A	E	I	O	U	B K S	C L T	D M V	F N W	G P X	H Q Y	J R Z
B C D		0										
F G H							X					
J K L												
M N P												
Q R S												
T V W					X			0				
X Y Z												
A												
E								0				
I												
O												
U										X		

Figure 4

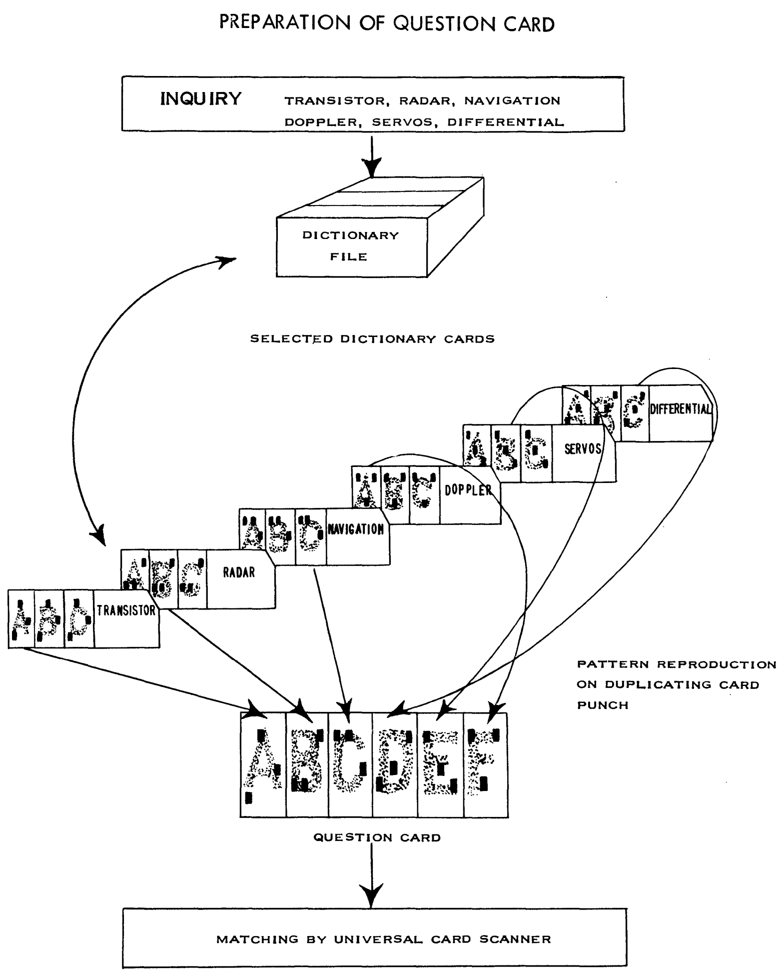

PREPARATION OF QUESTION CARD
INQUIRY
TRANSISTOR, RADAR, NAVIGATION
DOPPLER, SERVOS, DIFFERENTIAL
DICTIONARY
FILE
SELECTED DICTIONARY CARDS
TRANSISTOR
RADAR
NAVIGATION
DOPPLER
SERVOS
DIFFERENTIAL
PATTERN REPRODUCTION
ON DUPLICATING CARD
PUNCH
QUESTION CARD
MATCHING BY UNIVERSAL CARD SCANNER

Figure 5

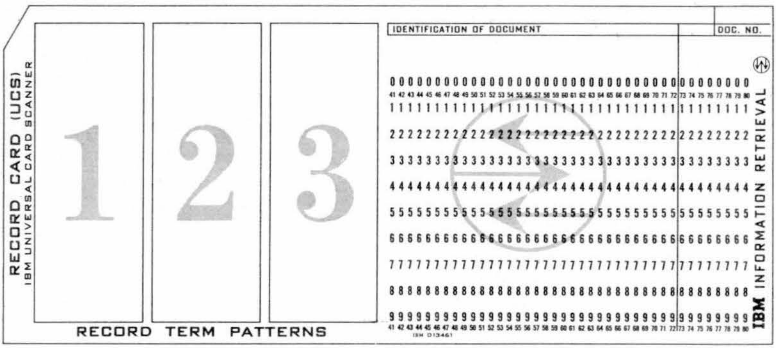


Figure 6 Record Card

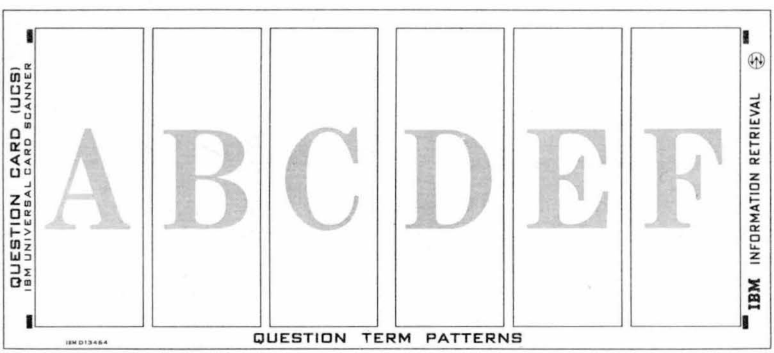


Figure 7 Question Card

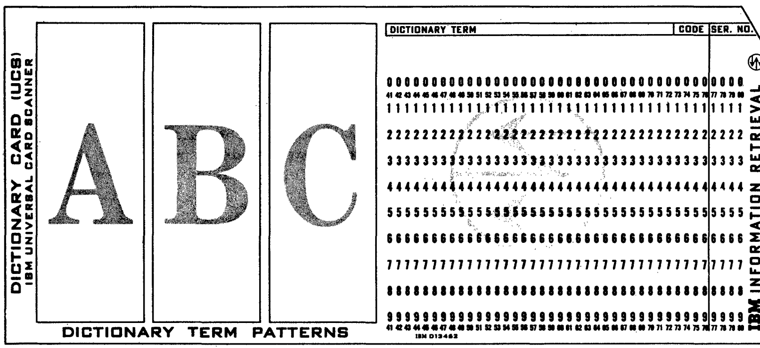


Figure 8 Dictionary Card, Front

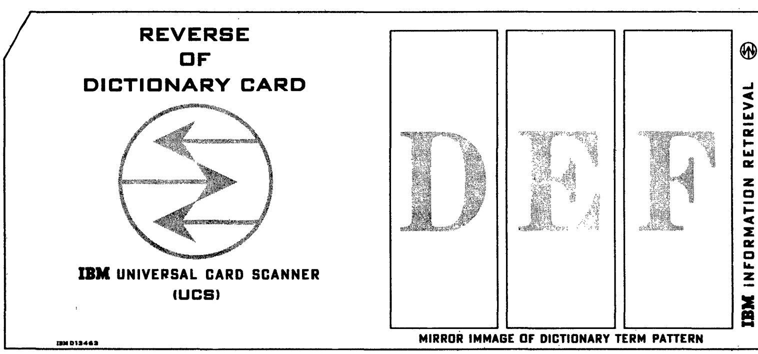


Figure 9 Dictionary Card, Reverse

Example 2: Bibliography Project on Information Retrieval and Machine Translation Literature

This experimental retrieval system was established by Mr. P. James at the Information Retrieval Research Department of the IBM Research Center to facilitate access to the sizable literature on the very subject of Information Retrieval and Machine Translation. The outstanding feature of this experiment is that the necessary encoding operations were carried out entirely automatically by an electronic data processing machine.

Preparation of Bibliographical Material

To start with, the basic information, collected from many sources, was manually punched into IBM cards in accordance with a standard format. This information, consisting of author, title and source of each of the documents involved, was individually recorded on sets of cards, i.e., a card or cards each for the author, title and source. This was the extent of manual preparation and the resulting master file of punched cards is intended to serve as means for generating any and all subsequent reference material required.

Auto-encoding of Bibliographical Material for Retrieval

The system provides for retrieval by means of keywords characterizing the titles of the documents. These keywords are compiled for each document by an IBM 704 machine which is programmed to analyze the words occurring in the titles. By means of table look-up, a predetermined set of insignificant or "common" words is excluded from the titles. The remaining words are considered to be significant and useful to serve as keywords for the retrieval operation. These words are then listed for each document and stored on magnetic tape.

Preparation of Record Cards for Mechanical Searching

In order to prepare record cards capable of being scanned by the Universal Card Scanner it is necessary to perform a number of transformations. Since at the instant of retrieval it is not known in which particular form a given word might have occurred in any of the titles of the documents, it is necessary to "normalize" varying word forms and to derive standard word stems which can be substituted for their variations wherever they occur. A routine has therefore been developed for normalizing words by machine in a systematic fashion according to a few simple rules and with the aid of a limited amount of table look-up.

A further requirement of the system is that four-letter code words be employed for creating code patterns in a 12 x 12 common pattern field on the record card. Rather than introducing a code word dictionary, these code words are derived by machine directly from the normalized words. In the present case this is accomplished by the method of "significant letter spelling" more fully described in the paper previously referred to.

There remains the problem of spelling the four-letter code words into the 12 x 12 matrix constituting the superimposed coding field of the record card. This too is done by machine by the method of chain spelling into a randomizing square in a fashion similar to the one described in connection with the previous example. In the present case the assignment of letters to the square is different from the previous one and has the form shown in Figure 10.

The three steps of transformation described above may be performed by the machine as a single operation. The result is a completed record card of the format shown in Figure 6 for each document. All code word patterns will have been punched into field 1 and the identification of the document and

its serial number will have been punched in the appropriate locations of the card.

Preparation of Dictionary Cards

In conjunction with the encoding operation a dictionary card is prepared by the machine. Its format is similar to the one described in the previous example. It is typical of the system that the dictionary is created as a result of the encoding operation. Consequently, the updating of the dictionary is completely automatic. The interfiling of new dictionary cards and the elimination of duplicates may be carried out mechanically by means of an IBM Collator.

Preparation of Question Card and Scanning of Record Card File

After an inquiry has been formulated, the preparation of an appropriate question card and the conditioning of the Universal Card Scanner are carried out in substantially the same manner as described in connection with the first example. This also applies to the scanning operation and the final print-out of the selected cards to produce a printed bibliography.

RANDOMIZING SQUARES

For code words derived by significant letter spelling

			E	T	A	O	N	I	S	R	H	L	D	C
			M	U	Y	F	G	W	K	P	V	B	J	X
			2	3	4	0	5	1	6	7	8	9	Z	Q
E	U	Q												
T	M	Z												
A	F	2												
O	Y	0												
N	W	3												
I	G	1												
S	P	4												
R	K	5												
H	B	6												
L	V	7												
D	X	8												
C	J	9												

Figure 10

CHAPTER VIII

RECENT IMPROVEMENTS IN TECHNIQUES FOR STORING AND RETRIEVING INFORMATION*

By J. C. Costello, Jr. and Eugene Wall**

This chapter is based on (1) a paper presented by the authors on January 13, 1959, to the Wilmington Chapter of the Society for Advancement of Management and (2) a paper ("New Horizons in Management--Information Systems") presented in Philadelphia by Eugene Wall on June 25, 1958 to a management symposium of the American Institute of Chemical Engineers and published in Chemical Engineering Progress (Jan., 1959, Vol. 55, No. 1).

In today's complex technological society, individuals and organizations must be provided at every responsibility level with the proper amount of accurate, up-to-date, pertinent information in the form required for each specific task, so that a more effective job can be done.

This is a broad field. It is not just "filing," which is the storing of physical documents. Here we are concerned with the total field of information handling, the adequate performance of all facets of which call for an effective

---

**Engineering Service Division, Engineering Department, E.I. du Pont de Nemours and Company, Wilmington, Delaware.

information system. The objectives of such a system would be to gain increased profits or savings by improving personnel effectiveness in relation to each specific task at hand. The information system must attain this objective, insofar as is economical, by efficiently providing, automatically or upon request, individuals and organizations with complete, pertinent, accurate, up-to-date information, in the proper quantity and form, routinely or nonroutinely correlated to the required degree of specificity.

In order to carry on such broad functions, six characteristics are required of an effective information system:

1. The system must effectively retrieve stored information--no matter what viewpoints and terminology are involved in originating and retrieving this information.

2. The system must, for the user, correlate information to the proper degree of specificity.

3. The system must contain a maximum of valuable information and a minimum of information with low or questionable value.

4. The system must provide access to expert human assistance.

5. The system must keep its stored information up-to-date and accurate.

6. The system must minimize paper flow but still meet the specific needs of users.

The first of these six characteristics, effective retrieval of information, is the heart of any information system. If information cannot be retrieved, there is little point in working on the other objectives and characteristics. Because

of space limitations, this paper will, therefore, concern itself only with effectiveness of retrieval.

There are two basic approaches to the problem of how to put away information and then to get it back again. The first approach is search harder for the desired information, which has been stored away with relatively few "handles" for retrieval. This is the approach which has been attempted for centuries and which has become increasingly ineffective as the volume and complexity of information have grown. The second approach is to increase effectiveness at the input side of the system--that is, we can provide more "handles" for retrieval, "handles" which may be useful from numerous viewpoints and with different technologies.*

This second approach--more effective indexing at the input side--should be carefully considered for large information systems. For such systems, the output-input ratio is very high. We have found examples in our work where, for every item entered into the collection, there are many items referred to by users. Hence, the place to expend effort is on the relatively low-volume input. Each document must be provided with index entries such that the user, no matter what his viewpoint or terminology, will have a satisfactorily high probability of retrieving it if the document is pertinent to his needs.

How can the indexer provide the needed index entries without knowing what viewpoints and terminology potential users may employ? We in the Engineering Department of the du Pont Company have concluded that the indexer should be

---

* The technical problems of terminology and viewpoint have been defined and described in detail in earlier papers. A concise discussion of these problems may be found in reference[1] of the bibliography.

provided with a "reminder list" (a thesaurus, in effect) of words generally associated with those words already used by the originator and by the indexer himself in connection with any given document.[2] This would be an extension of the concept of the "authority list," a device well known to librarians. If an author speaks of "distillation," the "distillation reminder list" should include such words as "rectification," "evaporation," "boiling," "separation," etc. The indexer may choose from the "list" those words which will be appropriate additional index entries for the document at hand. A descriptive name for such a crude thesaurus is "Word-association Matrix." This "tool" can also be employed at the retrieval end of the system to remind users of other ways of phrasing their questions, but this is advisable as a general practice only when the output-input ratio is low rather than high.

A loose but adequate basis for deciding if words are associated is whether or not they appear as index entries for the <u>same</u> document.[3] A Word-association Matrix can then be produced mechanically on computer or conventional punched-card tabulating equipment. With such techniques we have produced a Matrix from the past index entries of documents previously entered into our system. Thus, our Matrix includes the viewpoints and terminologies of earlier originators and indexers of information. It provides, for each term in the vocabulary of index entries, a list of associated terms, listed in order of frequency of association under the main term in question (Figure 1). This crude thesaurus can be used in its original form, it can be updated by adding subsequent indexing information, or it can serve as the basis for creating a more precise and complete thesaurus.[4] During indexing, the indexer first notes, as index entries, those pertinent terms which appear in the document being indexed. He then adds implied terms based upon his own knowledge; these will usually be synonyms and broader generic terms. He then refers to the Word-association Matrix (or to a refined

AIR POLLUTION

1600- 1

1600	AIR POLLUTION	50	
18600	CONTAMINATING - CONTAMINANTS - /SEE ALSO IMPURITIES/	50	100
1525	AIR /SEE ALSO ATMOSPHERES/	48	96
86425	WASTES - WASTE /SEE ALSO SCRAP/	39	78
35775	GASES - GASEOUS /SEE ALSO FLUIDS/	21	42
35150	FUMES	20	40
27425	DUSTING - DUSTS	16	32
75525	STACKS /SEE ALSO FLUES/	16	32
33675	FLY ASHES	15	30
4825	ASHES	13	26
38600	HEATING - HEATERS - HEATED	13	26
17375	COLLECTING - COLLECTORS	12	24
70400	SEPARATING - SEPARATIONS - SEPARATORS	12	24
8700	BOILING - BOILERS	8	16
17950	CONCENTRATING - CONCENTRATORS - CONCENTRATE - CONCENT	8	16
62375	POWER PLANTS - POWER HOUSES	8	16
725	ACIDS /INORGANIC/	7	14
25550	DISPOSING - DISPOSAL	7	14
65950	RECOVERING - RECOVERY	7	14
79875	TESTING - TESTS - TESTERS - TESTED	6	12
5025	ATMOSPHERES - ATMOSPHERIC /SEE ALSO AIR/	5	10
18700	CONTROLLING - CONTROLLERS - CONTROLS	5	10
28950	EMITTING - EMISSIVITY	5	10
28025	EFFLUENTS	5	10
31350	FANS	5	10
38375	HAZARDS - HAZARDOUS	5	10
49800	MEASURING	5	10
51650	MISTS	5	10
66150	REDUCING - REDUCERS /SIZE OR AMOUNT/	5	10
68850	SAFETY	5	10
72850	SMOKE	5	10
74075	SOLIDIFYING - SOLIDS	5	10
16525	CLEANING - CLEANERS - CLEAN	4	8
17450	COLOR - COLORS	4	8
20900	CYCLONES	4	8
25475	DISPERSING - DISPERSERS - DISPERSIONS	4	8
35050	FUEL - FUELS	4	8
57450	PARTICLES - PARTICULATE /SEE ALSO POWDERS/	4	8
62475	PRECIPITATING - PRECIPITATES - PRECIPITATED	4	8
66875	REMOVING	4	8
69725	SCRUBBING - SCRUBBERS	4	8
70575	SETTLING - SETTLEMENTS /SEE ALSO SEDIMENTATION/	4	8
78000	SULFUR DIOXIDE	4	8
85100	VAPORIZING, VAPORS, VAPORIZED /SEE ALSO EVAPORATING/	4	8
85350	VENTILATING - VENTING - VENTS	4	8
86325	WASHING - WASH - WASHABLE	4	8
86475	WATER	4	8
22875	DESIGNING - DESIGNS - DESIGN - TYPE	3	6
28650	ELECTROSTATIC	3	6

Figure 1

thesaurus, if one has been created). For <u>each</u> term under which he has <u>already</u> indexed the document, he checks the association list. From the list, the indexer selects additional terms appropriate in describing the document. It can be seen that the indexer must be a competent technical generalist.

Accordingly, the basic problems of terminology and viewpoint can be solved as rigorously or as lightly as one may be able to justify on an economic basis, because there are numerous degrees of freedom possible in using the Word-association Matrix. Some of these degrees of freedom include the extent and frequency of updating or of refining the Matrix, how extensively the indexer uses the Matrix, how lengthy one makes each Matrix list (i.e., does one cut off lower frequency associations?), etc. By making decisions on each of these variables, both the cost and effectiveness of the indexing operation can be controlled. Obviously, for a document with any given value to posterity, some optimum combination and use of these variables is best.

So much for development of the indexer's ability to solve problems in terminology and viewpoint. What <u>form</u> of index will best utilize this ability? There are three general types of indexing and retrieval principles which have been developed for storage and retrieval of information. The first of these is known as classification. A classification is an arrangement of information retrieval terms which <u>assembles</u> concepts into classes according to an <u>order of likeness</u> and <u>separates</u> them according to an <u>order of unlikeness.</u> It is apparent that as the body of collected information grows, it becomes necessary to institute more and more classes, sub-classes, sub-sub-classes, etc.

The second retrieval principle is conventional subject heading or alphabetical indexing, with which all are familiar.

The third retrieval principle is known as concept coordination (Figure 2). This is the generic name for various systems which analyze items (documents, drawings, etc.) into a set of terms or index entries, and provide for the retrieval of any particular item as the intersect of two or more terms. In such a system, any item (such as a person, a document, a machine, a process, etc.) is indexed under each of the individual concepts which describe it, not under combinations of concepts. The combinations are freely generated ("coordinated") by the searcher during retrieval. Thus, just as innumerable English words can be constructed from the 26 letters of the alphabet, a great number of items can be described in a concept coordination system with a relatively small vocabulary. Because permutations and combinations of concepts need not be indexed, the problem of system size is simplified.

There are two classes of concept coordination systems, distinguished by their diverse use of "system units" (Figure 3). "System unit" means an individual card, section of tape, or the like, which is manipulated during retrieval operations. The first class of concept coordination systems has one unit per document. This class includes edge-punched and most machine-punched card systems. The other class of systems has one unit per vocabulary term or concept and includes Batten card systems and coordinate indexes, which have significant advantages in index size, flexibility, and operating efficiency when large collections of documents must be handled. This is because--in such systems--only those concepts or units pertinent to the question need be searched, whereas the first-mentioned type of systems requires in principle the searching of the entire file. The criteria of choice between the two classes of systems have been described in more detail elsewhere.[5]

The physical size of an index system of the second class (or selective type) of concept coordination depends almost

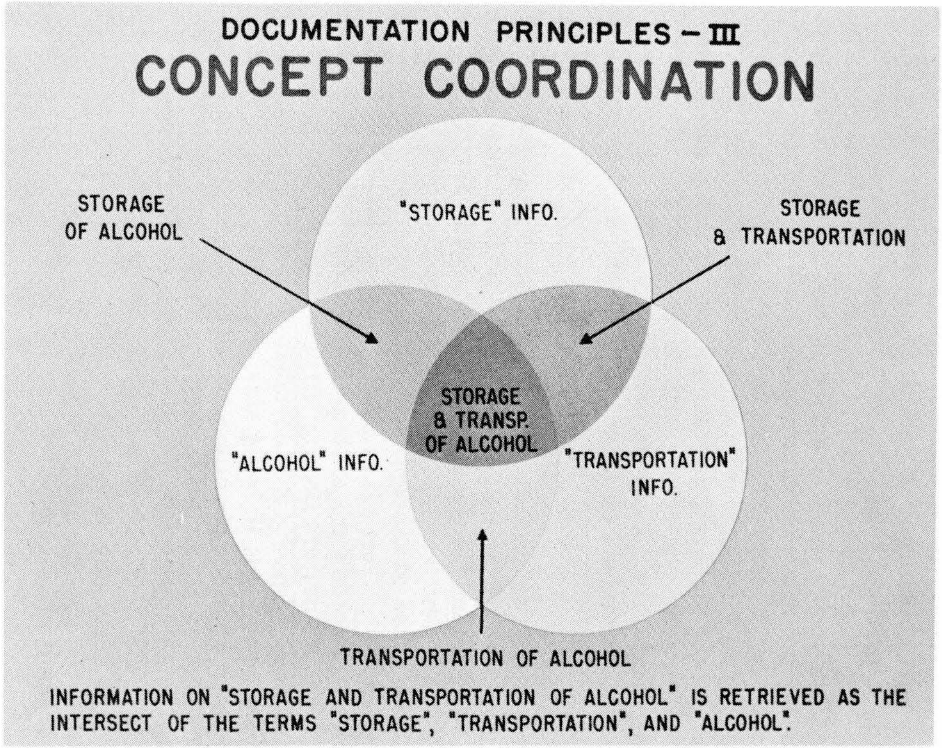

DOCUMENTATION PRINCIPLES — III
CONCEPT COORDINATION
STORAGE OF ALCOHOL
"STORAGE" INFO.
STORAGE & TRANSPORTATION
STORAGE & TRANSP. OF ALCOHOL
"ALCOHOL" INFO.
"TRANSPORTATION" INFO.
TRANSPORTATION OF ALCOHOL
INFORMATION ON "STORAGE AND TRANSPORTATION OF ALCOHOL" IS RETRIEVED AS THE INTERSECT OF THE TERMS "STORAGE", "TRANSPORTATION", AND "ALCOHOL".

Figure 2

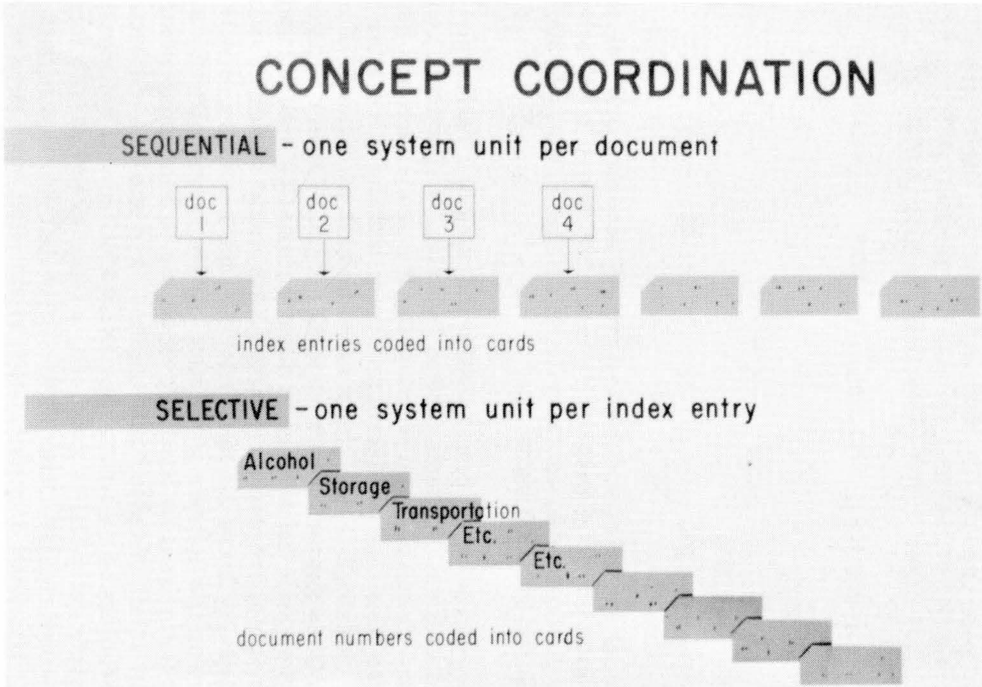

CONCEPT COORDINATION
SEQUENTIAL - one system unit per document
doc 1
doc 2
doc 3
doc 4
index entries coded into cards
SELECTIVE - one system unit per index entry
Alcohol
Storage
Transportation
Etc.
Etc.
document numbers coded into cards

Figure 3

completely upon the size of the vocabulary of indexing terms. Based empirically upon the experience to date of a number of operational concept coordination systems, the relation between total number of indexing terms, the vocabulary V, and D, the number of items in the collection, is approximately:

$$V = 4,200 \log_{10} (D + 700) - 11,600$$

and

$$\frac{dV}{dD} = \frac{1,800}{D + 700}$$

It can be seen that vocabulary growth becomes quite slow. We would expect a vocabulary to contain about 5,200 terms when the collection contains 10,000 items but only 9,300 terms for 100,000 items.

How do the three principles of indexing and retrieval--that is, classification, alphabetical indexing, and concept coordination--compare when considering the production and use of a Word-association Matrix and the providing of relatively deep indexing? First, let's examine classifications. Classifications gather narrow concepts together under broader concepts, but the subject in question must be entered under all known appropriate classes. This would result in an unmanageably large classification and the usual practice is to skimp on the number of entries. When large collections of documents must be handled, resulting in additional subordination within classes, sub-classes, and sub-sub-classes, the problem of categorization is intensified. If relatively complete retrieval is required, a widespread search through the entire classification is thus necessary to insure that all categorizations of the concepts in question have been located. This is not practical.

Classification is a suitable retrieval tool only when any of the following conditions prevails (1) the subject field to be indexed is narrow in its scope, (2) the classification will be used only by a small group of people who can learn it well and agree upon its categorizing conventions, or (3) the number of documents involved is relatively small.

When developing conventional alphabetical indexes, retrievability will be poor unless the subject at hand is indexed generically to the appropriate degree. For example, during retrieval of information on "liquid-liquid separation," the searcher may fail to look for data indexed under "distillation." The problem is intensified when combinations of concepts are involved. Should care be taken to index information under all possible headings, the index becomes unmanageably large. In practice, therefore, only a few main subjects are indexed.

Problems of viewpoint and terminology exist when using concept coordination to the same degree as when using conventional alphabetical indexing. Fortunately, when using concept coordination, the problems can be attacked and solved on the <u>individual concept</u> level rather than on a <u>combination of concepts</u> level or total system level. Production of a Word-association Matrix, or of a thesaurus, becomes a practical, feasible consideration rather than merely a theoretical one. Also, with concept coordination, deep indexing is practicable because of the relatively small system size.

The use of concept coordination does, however, bring about one peculiar problem of its own. This is the problem of syntactics, or word order. If syntax is disregarded, we may erroneously equate "cooling water" to "water cooling," or even "venetian blind" to "blind Venetian" (Figure 4). This is to say, if we look for the intersect of the concepts "blind" and "venetian," we may find <u>both</u> "venetian blind" and "blind Venetian" discussed in the documents to which we are referred. The problem of syntactics may be solved by setting up

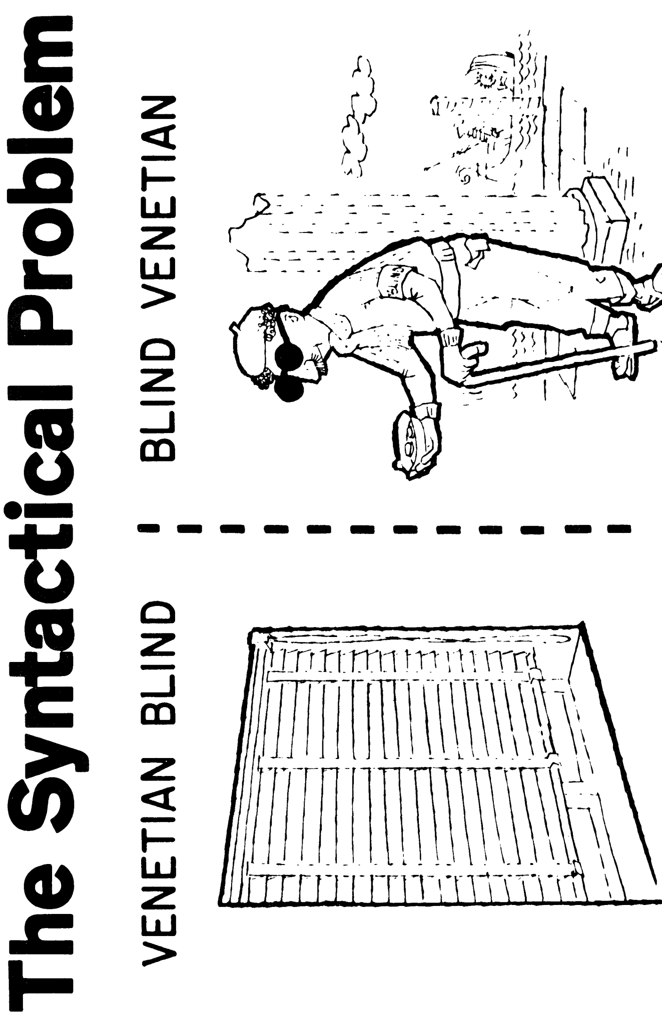


Figure 4

a new concept in the vocabulary (for example "blindness") or it may be solved by attaching role indicators to our index entries. For example, we can tag the concept "water" with a role indicator noting whether "water" is passive or active in the operation. Then, if we want "cooling water" (i. e., water for cooling), we will search only for those index entries which consider water as an active agent. Less than a dozen such role indicators, plus a few "association links" or "interfixes"[6] appear to be able adequately to handle the syntactical problem. These techniques will be discussed in more detail later.

The three principles--classification, alphabetical indexing, and concept coordination--may then be compared as shown in this table. The comparison is on the basis of equal retrievability or effectiveness (Figure 5).

In the Engineering Department, we began our work in this field by using concept coordination on technical reports. Indexing depth is approximately 20 index entries, or access points, per report. The index consists of a coordinate index, double-dictionary type book with identical pages in each of the two independent sides (Figure 6). When using the index for retrieving, the searcher chooses a combination of concepts expressing his "question," opens one side of the index to the word describing the most important of the concepts under question, and opens the other side of the index to the second most important concept. He then matches report numbers entered under each concept involved. The numbers which match will be those of reports pertinent to the combination of the two concepts searched. These matching numbers may then be matched with a third concept, etc. After obtaining the report numbers which are indicated to be pertinent to the question at hand, the searcher refers to a list of report titles, arranged in numerical order, to choose the most pertinent reports for further perusal. This title list includes under each title the index entries for that particular

Comparison of Documentation Principles

	Classification	Alphabetical Index	Concept Coordination
Cost	Very high	High	Low
Simplicity	Very poor	Poor	Very good
Adaptability	Very poor	Fair	Good
Compactness	Poor	Very poor	Good

This comparison based on equal retrievability.

Figure 5

MATERIALS OF CONSTRUCTION (8 - research on)

	121A		233D	64E		216B			219A

MATERIALS OF CONSTRUCTION (10 - design of)

			233D						

MATERIALS OF CONSTRUCTION (11 - acted upon)

90A		242A		24D 64E					9C 219A

MATHEMATICS (1 - use of)

		112B 112D 112F		184A 184B 184C 184D 234A	65G		107G 137A 137B	38A	229A 229B 229C 229D 229E 229F

MEASURING; MEASUREMENT (1 - use of)

			3A	124A 124B			67E		89B

MEASURING; MEASUREMENT (8 - research on)

50E 100F 100J 150A 190B	11C 61A 71C 221C	172B 202C 202D	3A 133A 133B 133C 133E 163C 163D 233F 253C	154A	25C 25E 45F 235A 235K 235N	136B 216E 266G	27E 27I 37F 57E 97E 197A 237H	78H 188C 188D 188E 188F 188G 188H 188I 188J 188K	39D 39G 89A 89K 89L 119D 119E 269D

MEASURING; MEASUREMENT (10 - design of)

			3B				67F		

MEASURING; MEASUREMENT (11 - acted upon)

	61A 61B						47G		89A

MECHANICAL PROPERTIES (1 - use of)

140A 190A 190B 190C 200B	1D		183A 263D	224A		186A			119A 159A

MECHANICAL PROPERTIES (2 - causes)

	211B	152B		134A		36D	27M		119B 179B

MECHANICAL PROPERTIES (8 - research on)

190A	1P 141A 161B	152D	53B 53C	44G 134A 204B	175C			178A 268A	119A 179A 189H 189I

MECHANICAL PROPERTIES (9 - effects)

10F 80B 80C 80D 190A 190B 200A 200C	1E 1M 31C 31G 201B	22J 92A 152B 162F	83B 93C 163A	124H 124I 134A 184A 194A 204C	135A	36C 36D 46A 46C 46E 46F 46G 46I 86B 206G	27B 27C 27F 27M 27N 27P 37B 97A 187A 187B 237E 237J 237L	78B 78D 78E 78F 178B 208B 268A 268D	29B 99A 119B 199A

CYCLOHEXANE (11 - acted upon)

	51A 241C 241I	212E 262A 262B 262C		214D	15A 15B 15D 15E 15F 25J 25K 25L 215H		17E 47G	88G 188D	9A 59A 59F 79J 259[illegible] 259[illegible] 259D

CYCLOHEXANOL (1 - use of)

					85F				

CYCLOHEXANOL (2 - causes)

		232G 252D	233H	64F 64G 164D	215B				

CYCLOHEXANOL (3 - reactants)

20A			113A 113B 113C				7F 7M	8A	

CYCLOHEXANOL (4 - special agents)

		252A			125A			87E	

CYCLOHEXANOL (5 - medium or solvent)

	191A		203E 203G	4C	85A 85E		7K	8F 138C 158A	

CYCLOHEXANOL (6 - contaminant)

	41G 191D 241C 241H		233F		85D 195A 215C		47G 227A 227B	258B	139D 259C

CYCLOHEXANOL (7 - reaction product)

						106C			

CYCLOHEXANOL (11 - acted upon)

	51A 111I 191D 241B 241C 241H	112C	33C 33J 233A 233B 233E 233F	84A 84C 164D 254C		76A 76B 76F 96B 106D 106F 106G 116I 256B 256C	7B 47D 127A 127B 127D 127G 127H 227A 257D 257F	258B	139D 259C

CYCLOHEXANOL-COLUMN (SEE CYOL-COLUMN)

CYCLOHEXANONE (6 - contaminant)

	241B				195A				

CYCLOHEXANONE (11 - acted upon)

	241B								

CYCLOHEXOXYMETHYL ACETATE (1 - use of)

				34A					

CYCLOHEXOXYMETHYL ACETATE (7 - reaction product)

				34B					

CYCLOHEXOXYMETHYL ACETATE (11 - acted upon)

				34K					

CYCLOHEXOXYMETHYL BENZOATE (1 - use of)

				34A					

Figure 6

document, thus providing a condensed abstract.

Our experience with this index has been very favorable. References to our technical reports have increased significantly over references which resulted from the earlier index, an alphabetical subject heading list. It is indicated that the retrievability of information from the new index is at least two to four times greater than for the old index and, of course, there is still plenty of room for improvement.

How may such an index be produced? The first requisite is to have indexable material. Any source of information can be considered indexable material. In the traditional viewpoint, a physical written record which has been prepared for the purpose of preserving and transmitting information for reuse is justifiably considered indexable material. Such records include books, periodicals, pamphlets, reviews, personnel records, accounting records, technical reports, research memoranda, test data, graphs, charts, design drawings, films, slides, and other similar types of records which depend on visual reception for successful transmission.

Collections of items may tend to vary widely with respect to similarity of characteristics such as format, organization of content, size and appearance. In this sort of situation, especially when the information content is quite non-homogeneous as concerns value per item, it is usually desirable (for economic reasons) to create a "standard item"--such as a summary sheet or the like--so that searchers may have common bases of reference by which to compare information content. One such standard item may summarize a number of minor, less valuable documents.

It is possible to index an item adequately by simply writing terms in one continuous list. Our experience in indexing has shown that it is often more desirable to provide for grouping the terms according to the general similarity of concepts to

which they refer. Six workable groupings of terms have been identified, and they are referred to as categories:

Category 1 - Names, such as plants, laboratories, sales offices, departments, geographical locations, project numbers, and other proper names or identifying numbers.

Category 2 - Chemicals which are identifiable by formula, such as compounds and elements; also families of elements and families of compounds.

Category 3 - Materials, mixtures, forms of energy, states of materials, names of products, etc.

Category 4 - Equipment, machinery, mechanisms, components, instruments, devices, meters, buildings, forms of materials.

Category 5 - Processes, operations, technologies, bodies of knowledge.

Category 6 - Attributes, conditions, concepts, forces, properties, characteristics.

Categorization of terms according to this system has been found to be useful in guiding indexers in their analytical techniques and to provide more familiar patterns of selection of terms by different indexers. Grouping of terms in categories greatly facilitated the production of the Word-association Matrix.

Using six categories for grouping indexing terms together is an attempt to establish a classification system for words. As in every known classification scheme, categories are not mutually exclusive, and there is unavoidable overlapping. However, Category 4 terms will always carry the connotation

of definable entities whereas Category 3 terms usually carry the connotation of quantity. Thus Category 4 terms are generally listed in the plural form, whereas Category 3 terms are entered as either singular or plural as necessary to indicate their essentially material nature.

Chemicals which are sold under a trade name are best indexed by their chemical names in Category 2 and by their trade names in quotation marks in Category 3. This same reasoning applies to pure ores and minerals - index the chemical name in Category 2 and the name of ore or mineral in Category 3.

Some Category 5 terms invite confusion unless handled with caution. For example, INSULATION may be properly a material, indexed in Category 3, and CLASSIFICATIONS may properly be devices in Category 4. Both INSULATION and CLASSIFICATION may properly be Category 5 terms. To avoid confusion, we have found it preferable to use the terms INSULATING and CLASSIFYING in Category 5, and generally in such situations, to use the -ing suffix.

In Category 6 (Attributes, etc.), terms are either adjectives, which qualify in some manner other terms in Categories 2, 3, 4 or 5, or they are properties, conditions, or characteristics of things, or forces which may act on things. Only rarely will Category 6 terms be entered as the plural form.

After a number of items have been indexed, the resulting coordinate index should be designed so that a searcher may successfully enter the index under any synonymous term (such as those above) and be directed to see that one specific term which has posted under it all the items which have been indexed under the referring synonyms or near-synonyms. For example, a searcher interested in items referring to the operation of chilling may find in the coordinate index

the reference CHILLING, SEE COOLING. Similarly, CLIPPING, SEE CUTTING; BOOKKEEPING, SEE ACCOUNTING: and BUFFING, SEE POLISHING. Where synonymous terms are alphabetically close to each other, generally there will be no "see reference." This applies to terms such as FILTERING and FILTRATION, POLYMERIZATION and POLYMERIZING, and CENTRIFUGATION and CENTRIFUGING. If a term in any category may be interpreted as having possibly more than one meaning, the term may be explained by a scope note in parenthesis, such as LEAD (Pb) or POWDER (EXPLOSIVE).

Now we come to the actual process of indexing. Indexing is essentially a four-stage intellectual process involving analysis, identification, evaluation, and description. This applies regardless of whether the system is built on concept coordination, classification, or alphabetical indexing of subject headings. The first three steps are necessary to guide the indexer to correct answers to the following two general types of questions:

(1) When this specific item was originated, what information did the originator consider valuable enough to record and transmit for future reuse? What did the originator intend to preserve for the benefit of others? What knowledge did the author want to make available to others -- and did he really do so?

(2) In developing the record to preserve and transmit the knowledge he considered to be of major importance, did the author record, in addition, any secondary or corollary information which, in this specific document, may be only supporting or background material, but which may have appreciable reuse value if evaluated from a different viewpoint?

Defined in another way, indexing is the process of providing proper terms to define concepts, after it has been determined: first, what the author intended to transmit in his item, and second, what other incidental information of reuse value was recorded. These two levels of analysis, identification, and evaluation generally take place during an initial general familiarization with an item and a subsequent closer, more searching examination.

First, the indexer will familiarize himself with the item he is about to index by studying the title and by reading those explanatory elements which may be available, such as (in the case of reports, for example) the abstract, foreword, summary, table of contents, conclusion, and list of appended material and illustrations. In some instances, the author will have provided a list of what he feels are appropriate terms to assist the indexer. At this point, prior to referring to the body of the report and the appended material, the technically trained indexer will be able to sense the intent of the author and the purpose of the item, and to identify the concepts treated in the item. Then, while the material reviewed is fresh in his mind, the indexer should write on the indexing sheet those terms which most concisely and accurately describe the concepts. The terms chosen will reflect the usage of words in the item and the word-usage habits of the indexer. The accuracy and adequacy of the terms selected will depend on the indexer's viewpoint and on his technical qualifications and competence.

Next, the indexer should examine the body of the item and appended material to determine two things: (1) is there additional information in the item which has not been adequately described by the indexing terms already recorded, and (2) have indexing terms been recorded which describe concepts treated only so briefly or summarily that, in fact, there is no information of value to anyone interested in references on those concepts? This second phase of analysis,

identification, and evaluation may result in the inclusion of additional indexing terms or it may result in the deletion of some initially recorded. It is in this second phase that indexers have their most important responsibilities - to identify valuable information not readily discernible in terms of the primary purposes of the item, and to evaluate for exclusion from indexing, information of negligible reuse value.

The development of the Word-association Matrix as a solution to the language problem in information storage and retrieval has been described earlier. Through the Word-association Matrix, there can be brought to the attention of an indexer the sum of system experience. Working term by term through his indexing sheet (as he has completed it to this point), each indexer, by consulting the Matrix, can consider individually the associated terms in each list to determine whether or not they are appropriate and pertinent for the more adequate and complete description of the information in the item being indexed. However, if the input-output ratio of the system is low, the use of the Matrix may be minimized at this time and used more extensively during retrieval.

After the proper terms have been selected to describe the concepts represented in an item, they are incorporated in a coordinate index along with the terms which describe all other items in the collection. The storage device may be a deck of Batten cards, a dual dictionary, a deck of machine-processed punched cards, or a magnetic tape. Location of desired information then consists of coordinating terms, regardless of the actual means of storage, to obtain logical products, (which means "all items dealing with this <u>and</u> this <u>and</u> that, etc."), logical sums (which means "all items dealing with this <u>and/or</u> this <u>and/or</u> that"), or logical differences (which means "all items dealing with this <u>but not</u> that").

Unfortunately, indexing and storage, as they have been thus far described, fail to provide the storage terms with those designators of relationship by means of which written and spoken language make sense. Sentences are made up of subjects, verbs, and objects, with appropriate modifiers, connectors, and qualifiers. Work order and relationship are extremely significant in sentence construction. The order of sentences with respect to each other is essential in order to transmit a group of concepts with intended meaning as a paragraph, as a chapter, as part of a larger document, or as a complete message.

Where collections of items are small, the absence of syntactical control elements is not particularly troublesome. In large collections, however, if provision for syntactical control elements is not made, the number of false associations obtained may be so large that the "noise" problem will cause a substantial decrease in system effectiveness. To cope with syntactical problems and to minimize false retrievals, "role indicators" and "association links" can be developed. These can be appended to terms in the indexing process to provide basic elements of grammar and sentence construction for greater specificity and selectivity in retrieval. This technique is similar to that described earlier by Whaley.[7]

The magnitude of "noise" in retrieval is dependent on the number of terms used to index a document and the degree to which they are bilaterally exclusively associated. Certainly not all the possible false drops will ever be retrieved, since questions involving some pairs of terms would never be asked, simply because of the unrelatedness of the concepts represented.

The potential amount of noise, that is, the relative number of potential false associations and false retrievals, can be reduced by appending to the item numbers, in the index-

ing operation, a "link" which binds together those terms which, when coordinated, will locate only real information. Linking of terms may be accomplished by affixing a letter to those terms. Links serve to accumulate into sentence-like association those indexing terms which describe the existence of information in an item on a concept or a number of related concepts. The use of links results in the reduction of the ratio of false retrievals to true retrievals. Links serve only to associate terms into a relationship which is more cohesive than if they were unlinked. In essence, their use results in an intellectual (not physical) sub-sectioning of the item. However, links provide no indication of the relationship which individual words within a linked group bear to each other. To provide this, a system of role indicators has been developed. These indicators are assigned to terms by indexers during the indexing operation.

There are three basic requirements which must be met by role indicators:

1. They must be indicative of broad concepts which are encountered very frequently in the particular environment of the information system.

2. They must, insofar as possible, be non-ambiguous among themselves (i.e., mutually exclusive) and--accordingly--

3. They must be few in number.

Because of the nature and degree of homogeneity of the information with which the Engineering Department is concerned, eleven roles (Figure 7) have been found to provide the additional exactness of indexing required to reduce to a satisfactory minimum the relative number of false retrievals obtained from typical coordinations. Roles are not

Roles

1. Uses of (for); Applications of (for); Used to (for)
2. Causes; Influences; Independent (Controllable) Variables
3. Reactant; Input; Raw Material
4. Special Agent
5. Medium; Vehicle; Solvent
6. By-product; Waste; Scrap; Contaminant
7. Product; Output; Manufacture, Production, Fabrication or Synthesis of
8. Research on; Development of; Investigation of
9. Dependent (Affected) Variable
10. Design of
11. Physically Processed, Treated, Changed, Handled, etc.; Passive

Figure 7

assigned to terms in Category 1 nor to adjectives or adverbs in Category 6.

A term with role assigned is essentially a precoordination of the term with an implied definitive concept term which imparts to the term-plus-role an element of syntax or word ordering so that stored information produces fewer false associations. In a coordinate index in which roles are used, items referring to packaging of "Mylar"* would be retrieved by coordinating appropriate term-plus-role for PACKAGING and "MYLAR" 11: PACKAGING and "MYLAR" 1 when coordinated would retrieve items referring to packaging <u>using</u> "Mylar."

Such a system of role indicators eliminates the need for storing the index terms in a fixed order, which is often done so that relationships among terms may be retained.[8] By using role indicators, in the fashion described above, it becomes possible to define inter-term relationships in an inverted, selective-type index and the expensive, time-consuming sequential search (which is necessary when using conventional systems) is avoided.

Although links and roles can reduce to a very acceptable minimum the amount of "noise" in retrieval, they can not entirely eliminate it. Indexers will find that the use of links and roles in indexing will necessitate a careful analytical approach to the information. Terms cannot be selected carelessly since they must be assigned appropriate links and roles. Further, the use of roles necessitates use of a set of conventions to insure their use in a consistent manner. With only a few roles, such conventions can be simple and few in number. As a result, the quality of indexing is significantly improved and this in turn favorably affects the effectiveness of retrieval.

---

* Du Pont's trademark for its polyester film.

After indexing has been completed for a predetermined number of items, there will have been accumulated a number of indexing sheets, one for each item indexed. These sheets will carry item numbers, indexing terms, and categories, links, and roles as appropriate. The completed indexing sheets are delivered to a keypunching unit so that there may be prepared one punched card for each combination of item number-plus-term-plus-role-plus-link. These cards have been keypunched so that two tabulations may be prepared for use in editing. One is referred to as the term-on-item tabulation, the other as the item-on-term tabulation.

The term-on-item tabulation is merely a recapitulation and rearrangement of the terms, roles, links, and categories as they appear on indexing sheets. Several copies of this tabulation are prepared for use in editing, so that editors can, as necessary, reconstitute the messages of the linked terms. This interpretation of terms is often necessary so that the exact sense in which a specific term was used may be known.

The item-on-term tabulation is a list in which terms have been ordered alphabetically with item numbers listed after them. In this tabulation, the item numbers carry the links with them. On the same line with each item number-plus-link, there is shown the role which was assigned to the term in that item. One copy of the item-on-term tabulation is the work copy, and the indexer who is responsible for directing and guiding editing activity uses this copy for recording decisions. Other editors, in a cooperative activity, consult dictionaries, handbooks, a Word-association Matrix (if one has been developed) and any available records of previous editing decisions.

Editing determines by consensus what changes, deletions, or additions should be made in the system's vocabulary so that optimum access is provided to the stored information.

This requires consideration of each individual indexing term, starting with the letter "A" and progressing through the entire vocabulary. The results of these deliberations will be to make various decisions and to mark the decisions on the item-on-term tabulation. These decisions are generally of the following types:

1. Which synonymous or nearly-synonymous terms should be added and provision made for cross-reference by designating "see references"?

2. Which synonym among several which have been used in indexing shall be the term on which item numbers will be posted, and hence, which other synonyms shall direct searchers to "see" the control synonyms?

3. Which nearly synonymous terms should be considered as having sufficiently different elements of meaning to justify searchers to "see also" the other near-synonyms, rather than to combine all item numbers under one term?

4. What the probable generic relationships are among terms, so that item numbers may be posted on terms of generically higher levels to provide for more rapid and efficient information retrieval by classes?

5. Which terms of two or more words should be broken down into two or more terms?

6. Which terms are so universally applicable to information or so nebulous in meaning that they contribute nothing to retrievability and hence can be deleted as indexing terms?

7. When scope notes should be appended to homographs to define meaning?

8. Which terms bear such clear and closely direct

relationship to one central concept that they may be grouped together as one term?

Ideally, much of the effort expended in editing may be eliminated by making available to experienced indexers, during indexing, a Word-association Matrix, which has been described earlier. Its use will call to the attention of indexers terms which are synonymous or nearly synonymous, terms which are on generically higher levels, and terms which are closely related to the central idea of specific terms as they are considered.

Editing (whether accomplished solely through committee deliberations, through use during indexing of a Word-association Matrix alone, or through some combination of the two) has as its objective the production of a coordinate index which will provide maximum access to stored information with maximum potential for retrievability.

The final form of the coordinate index may be a dual dictionary (Figure 6). Alternatively, the final form may be decks of term cards, or magnetic tapes. However, the form which results immediately from editing is a set of hand-posted term cards, one card for each term and role, listing the appropriate item numbers and appended links. This deck of cards and the information on them serve as the raw material for the preparation of coordinate indexes in whatever final form may be desired. In addition, the decisions of the editors are recorded on the backs of the term cards. Thus decisions are permanently available to successive editors.

References

1. Taube, M. and Wooster, H., Information Storage and Retrieval, Columbia University Press, New York, 1958, pp. 170-183.

2. Bernier, C. L. and Heumann, K. F., "Correlative Indexes III, Semantic Relations Among Semantemes--The Technical Thesaurus," American Documentation, 8, No. 2, April 1957, pp. 211-220.

3. Taube, M., Studies in Coordinate Indexing, Vol. II, Documentation Incorporated, Washington, D.C., 1956, pp. 72-80.

4. Luhn, H.P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information," I.B.M. Journal of Research and Development, 1, No. 4, Oct. 1957, pp. 309-317.

5. Taube - Wooster, Op. Cit.

6. Andrews, D.D. and Newman, Simon M., "Storage and Retrieval of Contents of Technical Literature," Research and Development Preliminary Report. Office of Research and Development, U.S. Patent Office, May 15, 1956.

7. Whaley, Fred R., "A Deep Index for Internal Technical Reports," Symposium on Systems for Information Retrieval, Western Reserve University, Cleveland, Ohio, April 15-17, 1957.

8. Perry, J. W., Kent, Allen, and Berry, M. M., Machine Literature Searching, Western Reserve University Press, Interscience Publishers, New York, 1956, pp. 91-99.

CHAPTER IX

THE MAGNACARD SYSTEM*

By Dr. R. M. Hayes**

A. General Description

Modern data handling requires the organizing, filing, and searching of great masses of data at high speed. Magnacard is a new concept in data processing, specially designed to solve these problems. It uses individual magnetic cards as the basic medium for storage of information. Machines have been developed by The Magnavox Company to handle the magnetic cards at high rates of speed - 6,000 cards per minute in typical sorting, selecting, merging, and file collating operations. These card handling units, when combined with a central electronic processing unit, form the Magnacard Processing System. Magnacard thus combines the high speed advantages of electronic processing and the high capacity of magnetic recording with the ease of handling inherent in the use of individual unit records.

Magnacard, thus, is a number of things: it is a magnetic storage technique; it is a storage medium; it is a system of operation. For each of these it is, in one sense, an extension of old well-established concepts. However, in another sense - because it represents a combination of these things -

---

* Reprinted with the permission of the Magnavox Company.

**Research Laboratory, The Magnavox Company, Los Angeles, California.

it is a very new concept.

For example, Magnacard is based on magnetic recording of information and, in this sense, is an extension of existing techniques of magnetic tape reading and writing. As such, it has all the advantages - in terms of storage density, erasability, information transfer rate - provided by this technique of information storage. As a result, Magnacard is so closely related to these existing techniques that it can directly replace magnetic tapes in virtually every usage. In another sense, however, Magnacard is a very new concept in magnetic storage. Present magnetic tapes are rapidly approaching a technological boundary on information rate. Although some tape equipment is now available which provides digital information rates of 60,000 characters per second, this equipment is expensive and probably is close to its technological limits. Magnacard, on the other hand, provides this information rate with its present techniques. In effect, the present Magnacard equipment represents a technological starting point rather than a technological upper limit. As further research is carried on, the density of recording can be increased, the speed of transport can be increased.

Magnacard is based on the concept of the card as the basic unit record storage medium. In this sense it is an extension of existing concepts of card storage. It is so closely related to these existing systems that it can directly replace punched cards or ledger cards used as file storage media. On the other hand, in at least two respects Magnacard is a new technique in card storage: first, it provides erasable storage in card form, so that posting and updating are simple; second, it provides large information storage capacity on each card so that complete data for a given item can be held on a single unit record.

Magnacard is based on handling equipment which provides high speed scanning of trays of cards combined with the

ability to choose alternative paths for card transport. In this sense Magnacard is an extension of existing card handling methods. As such it provides all the advantages - in terms of ability to sort, ability to merge and collate - afforded by card handling principles. However, in at least two respects Magnacard is a new technique in card handling: first, the speeds of operation - 100 cards per second - are greater than those of any other existing card equipment; second, the operation can be completely automatic, including multiple passes of cards, because the feeding and stacking stations are reversible.

Magnacard is based on filing techniques which provide rapid access to single items in large files of data. Two concepts of file mechanization are presently being worked on: one of these - the file block - is in a sense an extension of tape bin file systems; the other - the interrogation file - is similar to ordinary card files. However, in another sense, because of the combination of large capacity files, magnetic recording, and the individual card principle, these file mechanisms each represent more than merely an extension of existing techniques.

As it has been developed, Magnacard can be considered as essentially a component in data processing systems. In many respects, systems of operation suitable for Magnacard can be regarded simply as extensions of existing systems of data processing operation. When viewed in this light, Magnacard provides significant advantages in terms of information rate, speed of operation, data preservation, and cost. It can be used immediately as a direct replacement of magnetic tape and punched card files in existing data processing systems.

In other respects, however, Magnacard implies the capability for new systems of operation. The combination of advantages described previously permits not only the standard

systems of data processing but more complex combinations of card handling and internal processing. When these are used in conjunction with complete off-line card processing in file re-arrangement and card correlation, Magnacard provides unique answers to some highly specialized problems. The ability to handle files without re-writing, the high-speed card re-arrangement capability, the communication facility afforded by the transport means, all combine to produce a set of new capabilities.

B. Elements of the Equipment

The Magnetic Card

The Card: The basis of the Magnacard system is the use of individual magnetic cards for the storage of information. Measuring 1" x 3", the card consists of a Mylar base .005" thick with a .0005" iron oxide coating protected by a thin .0005" Mylar over-lay. The physical card itself represented a significant part of the Magnacard development. It has been engineered, after extensive research and testing, to withstand heavy usage under operating conditions. Two sources of wear are significant: surface friction due to continual contact between the card and the transport drum, and the forces exerted on the end of the card during stacking. Of these two sources of wear, the second is by far the more important, since the Mylar protective overlay protects the information bearing surface from the frictional forces. With respect to the force on the end of the card, specifications call for an operating life of 20,000 passes through the handling equipment. For files of cards subjected to daily processing, this will mean a useful life on the order of several years.

Data Reading & Recording: Information is recorded on the cards and read by techniques similar to those used with magnetic drums, using a sequence of magnetized spots recorded in tracks along the length of the card. Separate reading and

recording heads are provided, each consisting of twenty parallel tracks. Present recording uses the so-called "Manchester" representation and provides a density of 100 bits to the inch along the length of the card. On this basis each card has a capacity of approximately 5,000 bits. This is equivalent to 1,000 decimal digits or 600 alphanumerical characters. Since the technique of recording is magnetic, information can be added, erased, or changed as may be required by the processing.

Data Organization: The organization of this information capacity into characters, words, and other data groupings is a function of the particular central electronic processing unit and the requirements of the job. The magnetic card handling equipment is equally efficient with any of the possible data organizations, and Magnacard therefore can be compatible with any data structure required.

The Mechanical Structure

The mechanical handling equipment consists of a combination of a few basic elements - vacuum drums, feed-stack stations, transfer devices, hold stations, and reading and writing stations. These basic elements are combined to form the various handling units.

Vacuum Drum: The vacuum drums are the fundamental means for transporting cards. They consist of hollow drums about 8 inches in diameter and 1 inch high. Vacuum is applied continuously to the drum periphery through slots communicating with a hollow shaft. This vacuum provides a pressure differential between the outside and the inside of the drum. The difference in pressure holds the cards firmly on the periphery surface. The drum rotates at 12 revolutions per second, for a surface speed of 300" per second and a resulting maximum card rate of 100 cards per second.

Feed-Stack Station: Cards are successively fed into the drums and stacked from the drum by the feed-stack stations. The stations are dual purpose stations, capable of either function. The reversal of a station from one status to another is completely automatic, taking about .4 second. Feeding of cards can be either continuous at the maximum card rate, or intermittent with cards being released singly at rates up to the maximum card rate. Feed control is accomplished automatically, by pneumatic means using a high pressure vacuum controlled by a fast response electrodynamic valve. (This valve is one of the most significant developments of the Magnacard program since it provides the capability of very high speed control of air flow. Its use to actuate the feed control is just one of the places in which this device is important. Others are described below.) The stations are capable of accepting card magazines for storage of the cards. These magazines have a capacity of 3,000 cards each. Normally, they will be stored in file mechanisms with a ten-magazine capacity. The file mechanisms themselves are described in a later section.

Transfer Devices: These permit systems of drums and feed-stack stations to be designed with selective transfer of cards from one drum to another. These devices thus permit decisions to be made on cards and are the basis of the various sorting and collating operations. Cards are transferred from drum to drum by means of pneumatic jets controlled by the same type of high-speed electrodynamic valves. Selective transfer can be made at the free-running rate of 100 cards per second.

Hold Stations: These are provided for temporarily holding a card after it has been read, without removing it from the processing flow. This permits time for additional processing of the data before writing on the card; it allows cards to be merged from two separate feeding stations into a single stacking station; and it permits other cards to be transferred onto

the same drum for simultaneous circulation. The same pneumatic technique is used in the hold station as that used in the feed control.

Reading and Writing Heads: Reading of information from the magnetic cards and writing on the cards is performed by means of separate reading and writing heads, similar to those used in magnetic drum recording.

C. System Block Diagram

The system block diagram consists of eight basic blocks: (1) the mechanical unit; (2) mechanical control unit; (3) mechanical control register; (4) the card (in particular, card data format); (5) the read-write circuitry; (6) the buffer; (7) the buffer control register; (8) logical equipment. The system operation can be defined as follows: Information is read from the cards through the buffer to the logical equipment. On the basis of that information the logical equipment makes decisions concerning the cards and controls the operation of the mechanical unit by transmitting "program step numbers" to the mechanical control register. These in turn select specific program steps in the mechanical control unit which are wired to initiate the execution of functions in the mechanical unit.

To carry out this system operation, several lines of communication between the various units are required. They include the following: (1) control lines from the mechanical control unit to the various elements of the mechanical equipment; (2) control lines from the mechanical control register to select the required mechanical control unit program steps; (3) an information line from the logical equipment to the mechanical control register for transmission of the program step numbers; (4) an information line from the logical equipment to the buffer control register for transmitting buffer mode numbers and other control information; (5) control

lines from the buffer control register to the buffer; (6) information lines from the logical equipment to and from the buffer; (7) information lines from the buffer to and from the read-write heads; (8) information lines from the mechanical unit to the mechanical control unit to describe the status of the mechanical unit; (9) information lines from the mechanical control unit to the logical equipment to describe the status of the mechanical control unit and of the mechanical unit; (10) information lines from the mechanical control unit to the mechanical control register to describe the status of the mechanical unit; (11) information lines from the mechanical control register to the logical equipment to describe the status of the mechanical control register; (12) information lines from the buffer control register to the logical equipment to describe the status of the buffer control register; and (13) information lines from the buffer to the logical equipment to describe the status of the buffer.

D. Specifications on Mechanical Units

The following sections consider a limited number of possible arrays of Magnacard mechanical elements. These arrays are the ones which analysis over the past three and one-half years has shown will be the most generally useful. They involve the following individual devices: the one-drum mechanical unit, the two-drum mechanical unit, the four-drum mechanical unit, and the file unit. Their general characteristics are summarized as follows:

The one-drum unit in general will be a replacement for magnetic tape in Magnacard systems of operation. It can be used as the main input for file processing. It will also be useful in installations where the capabilities of the four-drum card handler are not necessary.

The two-drum unit is primarily useful where a matching of information rates is required. It will find its great-

est application where there is a requirement for large volume transcription of information from communication lines. For any other usage the machine is functionally too limited and can generally be replaced by the one-drum unit.

The four-drum mechanical unit is the most generally useful and represents the heart of the Magnacard system. This device, under the control of suitable logical equipment, can perform all of the routine card-handling operations, including sorting, merging, file searching, input, output, and transcription. Because of its general purpose capabilities, the four-drum device can well be used efficiently where any one of these functions might not consume its full operating time.

The file units are envisioned as being attached to any one of the other mechanical devices. They provide capability for automatic processing, with reasonable random access time, of large files of information. In their general usage they provide capabilities competitive with the tape bins. In a certain sense they also can be considered as competitive with the disc-type Ramac units. However, in a more important sense the two devices - the Magnacard file units and the disc-type Ramac - should be considered as complementary. Thus, where the Ramac type unit capabilities are required and economically justified, the Magnacard file unit cannot in any sense compete; where the processing can reasonably be performed on a sequential basis with a small amount of random access requirements, the Ramac units will be completely uneconomical and the Magnacard file unit will be at a competitive advantage.